MLTreeMap - Maximum Likelihood Placement of Environmental DNA Sequence Reads into Curated Reference Phylogenies

Dissertation zur Erlangung der
naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der
Mathematisch-naturwissenschaftlichen Fakultät
der
Universität Zürich

von
Manuel Stark
von
Erlen TG

Promotionskomitee:
Prof. Christian von Mering
(Vorsitz und Leiter der Dissertation)
Dr. Michael Baudis
Prof. Jakob Pernthaler
Dr. Alexandros Stamatakis
Dr. Thomas Wicker

Zürich 2011

Contents

Chapter 1

Zusammenfassung

Bei der Erforschung von mikrobiellen Gemeinschaften *in situ* stösst die traditionelle Mikrobiologie an ihre Grenzen, weil nur ein verschwindend kleiner Teil der Mikroben in Reinkultur gezüchtet werden kann. Die Idee diesen Engpass zu umgehen, indem man DNA direkt aus aus der Umwelt extrahiert und daraufhin sequenziert, führte zur Entstehung eines neuen Forschungsfeldes, der Metagenomik. Mit Hilfe des metagenomischen Ansatzes ist es möglich, objektive Informationen über alle in einer Probe präsenten Mikroben zu erhalten. Ein gewichtiger Nachteil ist allerdings, dass die gewonnenen Sequenzdaten nur fragmentiert vorliegen. MLTreeMap, das in dieser Dissertation vorgestellt wird, ist ein Softwarepaket, welches in der Metagenomik Anwendung findet. Es wurde entwickelt, um Einblicke in die phylogenetischen und funktionellen Eigenschaften von Metagenomen und den ihnen zugrundeliegenden mikrobiellen Gemeinschaften zu gewinnen. Hierfür werden die zur Diskussion stehenden DNA Sequenzen auf eine Reihe von relevanten Markergenen hin durchsucht und deren wahrscheinlichste phylogenetische Herkunft ermittelt. Zu diesen Genen gehören proteinkodierende phylogenetische Marker, 16S und 18S rRNA Gene und Marker für wichtige Stoffwechselwege. Beispiele für letztere sind die Gene der Schlüsselenzyme der Photosynthese, Stickstofffixierung, Methanfixierung und Ammoniakoxidation. MLTreeMap kann entweder direkt über das Web benutzt werden (http://mltreemap.org) oder aber auf einem lokalen Computer installiert werden. Wir veröffentlichten MLTreeMap im Jahr 2010 im Journal BMC Genomics [1].

Chapter 2

Abstract

Traditional microbiology has proven to be insufficient for studying entire microbial communities *in situ*, because only a small fraction of microbes can be grown in pure culture. The idea of circumventing this bottleneck by directly sequencing DNA from the environment led to a new field of research, called metagenomics. As a consequence of its approach, metagenomics provides a very unbiased view of all organisms contained in a sample, but it also has to cope with heavily fragmented sequence data. MLTreeMap, which is presented in this thesis, is a software framework designed to give insights into phylogenetic and functional properties of metagenomes and of the underlying microbial communities. It does so by detecting and phylotyping a series of relevant marker genes on the submitted DNA fragments. Among these genes are protein coding phylogenetic markers, 16S and 18S rRNA genes and markers for important functional pathways. Examples of the latter are genes coding for the key enzymes of photosynthesis, nitrogen fixation, methane fixation and ammonia oxidation. MLTreeMap is available as a web-server at http://mltreemap.org and also as a stand-alone version. It has been published in BMC Genomics in 2010 [1].

Chapter 3

Introduction

3.1 From taxonomy to phylogenetics

Modern biological taxonomy began with Carl von Linné (1707-1778). He hierarchically grouped organisms into species, genus, order and class. Further taxonomic ranks have been added since then, but in its essentials Linnés system is still in use today. Linné also invented the binomial nomenclature, which requires a species name to consist of two words, the genus name followed by a specific epitheton. He first applied his nomenclature to plants, the work on which was published in 1753 [2]. Some years later, he adopted it for the naming of animals in the 10^{th} issue of his book *systema naturæ* (1758-1759) [3, 4]. Even though his classification system agrees well with the idea of evolution, it is noteworthy that Linné still believed in the creation of life as described in the Book of Genesis. For him, the observed biodiversity was static. It took some years until paleontology, founded by Georges Cuvier (1769-1832), and its fossil records of earlier life on earth, gave rise to the idea of evolution [5]. Jean-Baptiste Lamarck (1744-1829) was the most famous predecessor of Darwin. He presumed that species change, due to environmental influences. But in contrast to what was later often said of him, he never claimed that these changes are inflicted by direct induction of the environment [6]. One of the main differences to our current view of evolution is that he believed it to be directed towards complexity and perfection. Lamarck was aware that the high prevalence of lower organisms stood in contrast to this belief, because

according to it most of them should have evolved to higher organisms a long time ago. Thus he further assumed spontaneous generation of new microbes. Today he is mostly known for his long rejected idea that acquired traits can be inherited. It is not without irony, that recent discoveries in epigenetics [7–9] show that his concept was not as wrong as it might seem [10].

There is a number of important achievements in science, such as the development of the periodic table [11, 12] or the establishment of the Hardy-Weinberg law [13], which have been reached simultaneously and independently by different researchers. Similar to this, Charles Darwin (1809-1882) was not the only scientist to come up with the idea of evolution and natural selection. But he was the first one to work it out in a very thorough and extensively documented way. More than twenty years passed between his first drawing of an evolutionary tree entitled with 'I think' in 1837 [14] and the publication of his famous book 'On the Origin of Species' [15]. With other scientists accepting his concept, the stage was set for systematics, the study of biodiversity and phylogenetic relationships between organisms. Taxonomy became a part of systematics, with an additional goal of reflecting the phylogeny of organisms in their nomenclature.

Phylogenetic reconstruction in early systematics was mainly based on morphological traits. In contrast to fossil organisms, where this constraint applies until today, molecular data nowadays provide a vast amount of additional information for living taxa [16,17]. Both reconstruction approaches, morphological and molecular, should be seen as complementary to each other, even though conflicting results sometimes led to tensions among taxonomists [18]. In the case of microbes, systematics proved to be especially challenging, because neither the traditional species concepts nor morphological traits are of much significance there [19, 20]. Thus it is commonly agreed that molecular phylogenetics is the best available method for microbial systematics. DNA similarity was defined as a criterion for establishing microbial species [21], also called molecular operational taxonomic units (MOTUs) [22]. The generally accepted 'species' threshold lies at a DNA-DNA hybridization value of 70% [21], resulting in average nucleotide identities of over 95% [23–25]. Considering its great importance for the taxonomy of all living organisms and its near monopoly for microbes, it is obvious that molecular phylogenetics has

become an integral and indispensable part of modern systematics.

3.2 Molecular phylogenetics

In the early stages of molecular phylogenetics, its supporters believed it to be a much more direct and objective method than the morphology based approach [26]. For the latter, external expertise is essential. Scientists have to judge which morphological traits are relevant and how they are to be weighted in respect to others. Molecular phylogenetics on the contrary was supposed to be based on pure statistical analysis. Margoliash, an important supporter of this idea, started conducting phylogenetic studies using Cytochrome c in 1963 [27, 28]. The field progressed further even though it became clear that there was an increasing amount of methodological problems, which did not conform with the ideal of a completely unbiased and unsupervised approach. It was Carl R. Woese and his group who first suggested to organize all living organisms into three domains of life: Bacteria, Archaea and Eukarya. Their analysis was based on small subunit rRNA (SSU rRNA) molecules, which were chosen because they occur in all self replicating systems, are evolutionary stable and easily isolated [29]. In the years to come, the SSU rRNA became one of the most important phylogenetic markers and the tree-of-life, which Woese et al. constructed in 1990, made a lasting impact [30] (Figure 3.1).

Some obstacles in molecular phylogenetics have been anticipated early, such as determining the number of mutations at a specific site, or the analysis of organisms that are either very closely or only very distantly related. But it was not expected that methodological problems should become as important as biological ones [31]. Already the first step of a phylogenetic analysis, the sequence alignment, leads to problems, which remain critical to this day. Needlemann and Wunsch developed a dynamic alignment algorithm as early as in 1970 [32], but even in the nineties, alignments by eyesight were still reported to be popular with some researchers [33]. Notwithstanding the arguments, which can be brought up in support of manual alignments, the huge amount of sequence data generated today, makes a computer based approach all but necessary. Another important advantage of alignment algorithms is that they are more objective and reproducible. Unfortunately, they

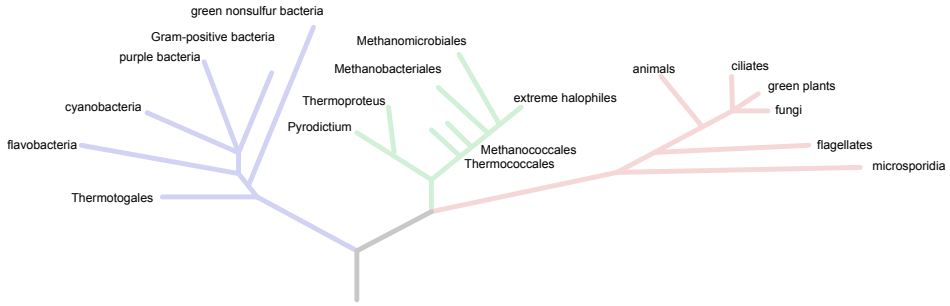


Figure 3.1: Schematic view of the tree by Woese et al. The tree was the first to contain data from all three domains of life: Bacteria (blue), Achaea (green) and Eukarya (red). Adapted from Woese et al. 1990 [30].

have their disadvantages too: Global alignments are suitable for closely related sequences, while local alignments work better with distant ones. There are attempts to combine both methods in so called 'glocal' alignment algorithms [34], but until now it is still up to the researcher to decide, which method works best for his data. Herein lies a potential of circularity, for this decision is likely to be based on information available only after the analysis. The insertion of gaps into alignments, which is unavoidable, if more than just very similar sequences are to be aligned, has its risks too: The reason for this is that longer sequences are very likely to contain identical passages just by chance. Excessive usage of gaps will bring these stretches together and thus create false positive relations, which have no biological significance [35]. Penalties are assigned to gaps and mismatches in order to alleviate this problem, but the choice of criteria for weighting them is again a matter of discussion. Depending on the point of view, some scientists favor the usage of subjective weightings, because they rate the benefits higher than possible disadvantages [36], while others still prefer a purely statistical approach [37]. To refine the weighting of mismatches in protein alignments, substitution matrices have been developed. They contain probabilities for all possible amino acid substitutions. Both, the PAM [38] and the BLOSUM [39]

matrices, have been constructed by analyzing several sets of reference alignments. As a consequence they are not universal. Instead the appropriate substitution table (e.g. BLOSUM62, BLOSUM80 etc.) has to be chosen according to the overall similarity of the sequences, which are to be aligned. Here we encounter the same problem as above, namely that the degree of similarity is most likely not yet known when the decision has to be made. While a pairwise alignment can be computed within a second, the computational costs increase exponentially with the number of additional sequences in a multiple alignment [40, 41]. Attempts to reduce these costs led to the development of a series of heuristic approaches [42–45], which of course are not guaranteed to find the mathematically optimal solution. Thus the quality of the input data is of even greater importance than in the case of pairwise algorithms. This is also reflected by the fact that most of the multiple sequence alignment algorithms require sequences suitable for global alignments (i.e. closely related sequences of roughly similar lengths). The algorithms are often guided by reference phylogenies, which are calculated at the initial steps of the alignment process. Obviously this is another case of circularity, because the resulting alignments will then again serve as input for the tree building programs.

The next step after sequence alignment is the construction of phylogenetic trees. Early tree building algorithms such as the one by Fitch and Margoliash [28] were based on simple distance matrices. Newer algorithms now mostly rely on either maximum parsimony [46] or maximum likelihood [47,48] approaches. Maximum parsimony methods assume that the reconstructed phylogeny with the lowest number of required evolutionary steps (i.e. the most parsimonious one), is the one which best reflects reality. An indisputable advantage of this method lies in its comparatively low computational costs. Unfortunately there are often several trees which maximize parsimony, and any decision between them is again dependent on external knowledge. A second important drawback of maximum parsimony is that it is prone to an artifact termed 'long branch attraction' [31, 49, 50] (Figure 3.2). The maximum likelihood approach on the other hand has higher computational costs, but is also assumed to be more reliable and accurate [51–53]. Maximum likelihood algorithms almost always yield only a single best scoring tree. In

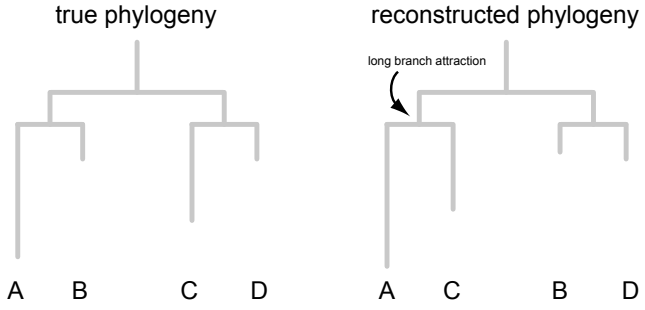


Figure 3.2: Long branch attraction. In a given phylogeny, those sequences with high substitution rates will be attracted to each other, independent of the actual relationships.

this respect, they get closer to the ideal of an unsupervised approach than maximum parsimony. Nevertheless they are not entirely free of human expert input either, for they need an explicit model of evolution and the choice of this model is not trivial [54]. In case of both, maximum parsimony and maximum likelihood, there is of course no guarantee that the 'best' tree corresponds to the true course of evolution. Due to the difficulties in choosing between models and methods many scientists tend to use the available software packages without going into the technical and conceptual details, as Suárez-Díaz and Anaya-Muñoz noted in their review of the topic [31]. Results from different methods are not necessarily weighted against each other, instead they are often displayed together and emphasis is put on the statements in which they agree (e.g. ref. [55–58]). Support for using human expertise, as opposed to pure statistical correctness, also came from Brinkman et al. [51]. They argued that a scientist, who is familiar with the data in his field, is often able to recognize which sequences will not fit into an alignment. Pruning these sequences from the dataset will enhance the analysis rather than invalidating it. MLTreeMap, which is the focus of this thesis, goes into this direction. It provides manually curated reference data and several analysis algorithms (e.g. maximum likelihood, maximum parsimony and the possibility of bootstrapping for both). Thus the user of MLTreeMap is able to conduct a comparison

of maximum likelihood and maximum parsimony algorithms with the same software tool and does not have to deal with many of the aforementioned problems.

The last point that I would like to discuss here concerns some criticisms of phylogenetics itself. If organisms pass on their genetic material solely from parent to offspring, the reconstruction of their phylogeny is possible to a large extent, albeit difficult as we have seen. But it is known since the studies of Avery et al. on pneumococci bacteria in 1944, that organisms can also exchange genes via horizontal gene transfer (HGT) [59]. The crucial question for systematics is, to what extent HGT takes place and whether it can adversely affect phylogenetic reconstruction. Schwarz and Dayhoff assumed in 1978 that HGT is not an important issue and claimed that basic metabolic genes have not been transferred horizontally [60]. This view has been challenged in the nineties by scientists, who postulated a web-of-life instead of a tree-of-life [61–64]. The debate is not over yet, for other researchers rose in defense of the tree-of-life [65, 66], while the rate of HGT is still under discussion [67]. Most likely, the extent of HGT is not equal across the various gene families and cellular functions, and also not across the various parts of a cell's genetic material (e.g. plasmids vs. chromosomes) [68]. Independent of what the result of this dispute will be, it already has direct consequences in the field of phylogenetic reconstruction. Scientists working on marker gene based algorithms have to take HGT into account and choose their markers accordingly [69].

3.3 Microbiology

Microbiology has been an expanding field of research since its beginnings in the late 17^{th} century. It has given insights into a largely hidden ecosystem, the importance of which cannot be overestimated. Microbes make up over one third of biomass on earth [70] and they interact closely with all other lifeforms, be it as symbionts or pathogens. The first scientist ever to see bacteria, was Antoni van Leeuwenhoek (1632-1723) with his homemade microscopes in 1676 [71]. After his discoveries, progress in microbiology was slow until the mid 19^{th} century. It was Louis Pasteur (1822-1895) in 1864,

who first demonstrated that microorganisms are not generated spontaneously - a discovery which led to the development of sterilization methods, which are essential until today. Another very important impact on the emerging field of microbiology was made by Robert Koch (1843-1910), who was the first to produce experimental evidence for the theory that microbes are responsible for many diseases. He introduced three criteria (known as Koch's postulates), which have to be met if a specific microorganism is to be established as the cause of a disease [72]:

1. The pathogen has to be present in each individual case of the disease under discussion.

2. The pathogen does not appear in other diseases as accidental non-pathogenic guest.

3. After being isolated and grown in pure culture, the pathogen can reproduce the disease in previously healthy individuals.

While this list had to be modified since then, because not all pathogens fulfill all criteria (e.g. viruses cannot be grown in culture without hosts), it enabled Koch and his followers to track down the causes of many human and animal diseases. Not least due to the third of Koch's postulates, the importance of pure culture increased tremendously, and existing culturing methods were improved and new ones developed. It was Martinus Beijerinck (1851-1931), who started, using the enrichment culture technique he had invented, to isolate and grow pure cultures of microbes from soil and water. The results of his research were a major inspiration for Baas Beckings famous tenet 'everything is everywhere, but the environment selects' [73]. The methods and principles established by those scientists and their followers provided a vast body of knowledge and paved the way for modern medicine. When in 1977 the first genome was sequenced by Sanger et al. [74], genomics has appeared as a new field of research. Yet until very recently, whole genome sequencing of microbes has been restricted to those organisms, which are cultivable - and estimates indicate that they represent only a very small fraction of the microbial world [75]. This results in an important bias of traditional microbiology, because the cultivable microorganisms are not necessarily representative of the

whole community [76]. A second important bias arises from the frequently quite anthropocentric research interests, which focus on human pathogens or microorganisms providing useful services [77]. Microbiology has often been said to be a 'method-limited' science [78], but several recent developments have greatly facilitated the transition from the study of model organisms in the lab to environmental studies. Apart from environmental genomics (also called metagenomics), which will be discussed in more detail below, other culture independent techniques have been developed in order to extend the limits of microbial research. Community proteomics for example transferred the metagenomic approach to the protein world [79–81]. Fluorescent *in situ* hybridization (FISH) [82, 83], which was invented in 1980 for detecting specific strands of DNA or RNA, was also adapted for environmental microbiology [84–86]. The combination of FISH and secondary ion mass spectrometry (SIMS) [87] (later improved to nanoSIMS [88, 89]) allows for simultaneous detection of a cell's identity and its functionalities. With this research ongoing, our knowledge on the microbial world is steadily growing and a new frontier lies ahead already: There is an increasing number of indications that extraterrestial bodies such as the planet Mars, or moons like Europa, Titan and Encelaudus, might harbor microbial life as well [90]. Examining these life forms if found, will be the next great challenge for microbiology and will have a lasting impact on our view on life and evolution as such.

3.4 Metagenomics

Metagenomics has emerged as a powerful new branch of science, overcoming the aforementioned biases of traditional microbiology by focusing on entire microbial communities *in situ*, including their many uncultivable members. The field was pioneered by Pace et al. [91, 92] , who first developed the idea of sampling DNA directly from the environment and thus circumventing the bottleneck of cultivation. The term 'metagenome' was coined some years later by Handelsman et al. for the collective genome of soil microorganisms [93]. It was soon adopted by others working on similar projects [94], and in time it became the name for this new branch of research. The interest in metagenomics has been steadily growing since then, so that in September 2009 a total of

200 metagenomic projects was registered at the Genomes On Line Database (GOLD) [95]. Even though other culture independent methods, such as single cell sequencing [96], are under development, metagenomics is very likely to keep the competitive edge in terms of high throughput for the time being. This is a central aspect because high throughput is essential if microbial communities are to be analyzed in their entirety. In a typical metagenomic workflow, DNA is obtained directly from a desired environment. Afterwards, the DNA fragments are either first cloned into bacterial vectors [93, 97] or directly shotgun sequenced [98] (Figure 3.3). Any following assembly of the sequence reads is usually limited or even impossible depending on sample complexity [99]. Several characteristics of the underlying microbial community can already be deduced from this fragmented sequence data, such as predicted genome sizes [100, 101], recombination rates [102] or certain functional properties [103]. Yet another important task, if one is to characterize and understand the community, is to determine its phylogenetic composition [104]. Due to the fragmented state of metagenomic data, this task is far from trivial.

There are two main classes of approaches to solve this problem: The first is 'unsupervised', which means that it does not require external reference information derived from fully sequenced genomes. Algorithms belonging to this class learn to bin the sequences to taxonomic groups based on intrinsic features, such as nucleotide usage patterns [105–107]. The (G+C)-content of DNA sequences is an intuitive example for this. Binning of metagenomic data based on this criterion, however, has been shown to be inferior to binning based on tetranucleotide patterns [108]. As a consequence of this result, Teeling et al. have developed TETRA, which detects and scores the frequencies of tetranucleotides [105]. Nevertheless the authors themselves state that TETRA works best for relatively long sequences (40 kb) and is not suited for sequences shorter than 1 kb. TETRA further encounters problems with the analysis of high complexity samples [105]. PhyloPythia [107] follows a similar approach, as it uses the oligonucleotide composition of DNA samples as binning criterion. Depending on sample complexity, 5-mers or more complex 6-mers of consecutive nucleotides have been shown to provide best results [107]. Similar to TETRA, PhyloPythia needs sequences longer than

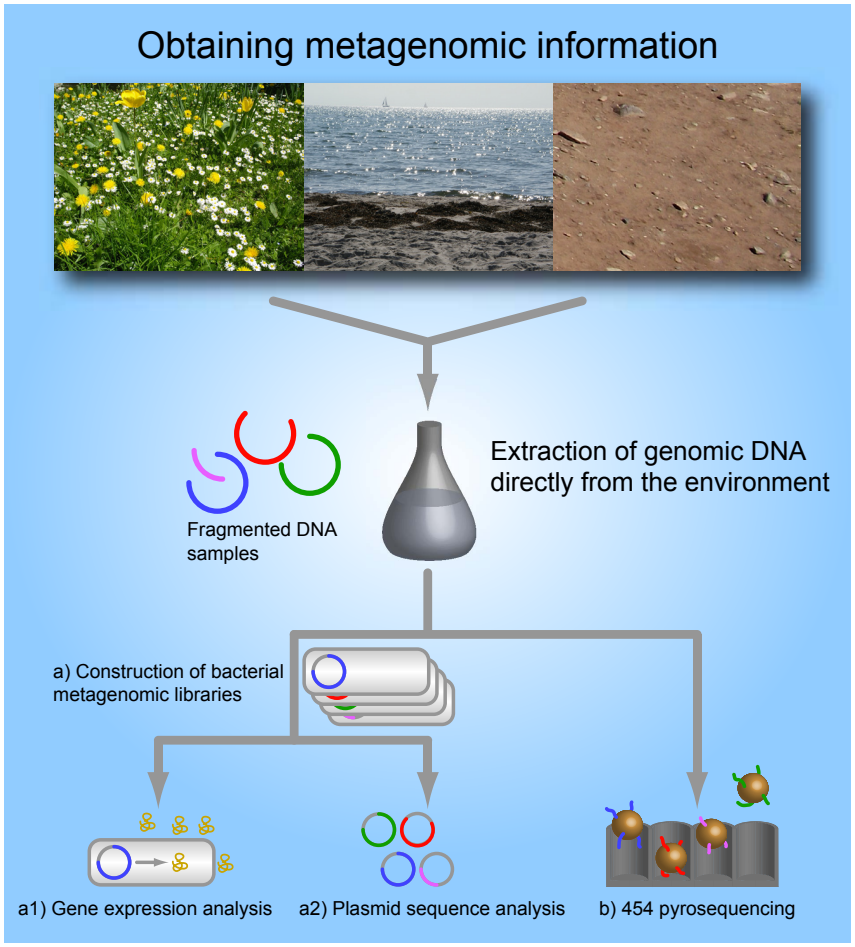


Figure 3.3: A typical workflow in metagenomics. DNA is isolated directly from the environment, resulting in sequence fragments of variable length. These fragments are then either shotgun sequenced, usually through 454 pyrosequencing (b), or first cloned into suitable vectors and transformed into host bacteria (a). Bacterial metagenomic libraries can either be used for gene expression analysis (a1) or for genomic analysis after plasmid sequencing (a2). Adapted from Handelsman 2004 [97].

1 kb to produce reliable results, which limits the value of both methods because the length of 454 pyrosequencing reads [109] still ranges only at about 400-500 bp [110]. The algorithm developed by Sandberg et al. makes use of a naïve Bayesian classifier for analyzing motif frequency distributions and is supposed to handle sequences shorter than 1 kb as well as longer ones [106]. Other algorithms belonging to the 'unsupervised' class are based on self organizing maps (SOM) [111], interpolated Markov models [112] or sequence similarity metrics [113]. Bayesian classifiers [106, 114] or the classifier used by PhyloPythia [107] are sometimes not counted among the 'unsupervised' approaches, because they can be trained with known sequences [115]. We do not follow this argumentation because providing such an external reference is not mandatory. The second class of approaches is 'supervised', meaning that it makes use of extensive reference data. A straightforward realization of this approach is just to determine the lowest common ancestor (LCA) of the best BLAST hits of a particular sequence, as implemented in the MEGAN software package [116]. Prior to an analysis with MEGAN, the user has to provide BLAST results for all query sequences. MEGAN then does the binning of the sequences guided by a set of parameters that determine the number and the minimum quality of the BLAST hits, which are to be taken into account for detecting the LCA. The ensuing results are visualized by a very comprehensive graphical user interface. MEGAN can be applied to entire metagenomes, but also to restricted sets of specific markers, such as 16S and 18S rRNA genes. While these two genes are still the most popular phylogenetic markers [97, 117], there is also a group of algorithms dedicated to phylotyping metagenomic datasets based on protein-coding marker genes [118–120]. ML-TreeMap, which is the main focus of this thesis, belongs to the 'supervised' approach as well, because it uses a series of marker genes as a reference to infer the species composition and selected metabolic properties of the query datasets.

Both approaches, 'supervised' and 'unsupervised' have their advantages and drawbacks: 'Unsupervised' methods are better in clustering unknown taxa, which may have only very distant relatives in the current reference phylogenies, whereas 'supervised' methods are superior in detecting organisms with low abundances in the samples [115]. Summarizing this part of the intro-

duction we can state that metagenomics provides a very broad and unbiased view of the microbial world, but the fragmented sequence data also impose new challenges for computional biology [121].

Chapter 4

The MLTreeMap algorithm

4.1 MLTreeMap - a short description

MLTreeMap is a software framework based on maximum likelihood that is designed to give insights into phylogenetic composition and functional properties of microbial communities. Phylogenetic analysis of MLTreeMap relies on a manually curated set of protein-coding marker genes, as well as on 16S and 18S rRNA data, according to which the query sequences are placed within externally provided tree-of-life phylogenies. MLTreeMap further searches the query sequences for the key genes of fundamental functional pathways, such as nitrogen fixation, photosynthesis and others. These genes, if present, are then placed into the respective gene family phylogenies.

A run of MLTreeMap starts with a BLAST search for the marker genes mentioned above. To prevent false positive hits, deep paralogs of the markers are searched for as well and discarded if found. Identified marker genes are then translated to protein sequences using Genewise [122]. The next steps include alignment [123], concatenation and gap removal [124]. After that, the sequences are phylotyped using RAxML [125], which employs full maximum likelihood (Figure 4.1).

A preliminary version of MLTreeMap has been outlined by von Mering et al. in 2007 [119] and already provided valuable results [126, 127]. In contrast to the new version presented in this thesis, the preliminary one was only available online, contained no functional markers and could only process one

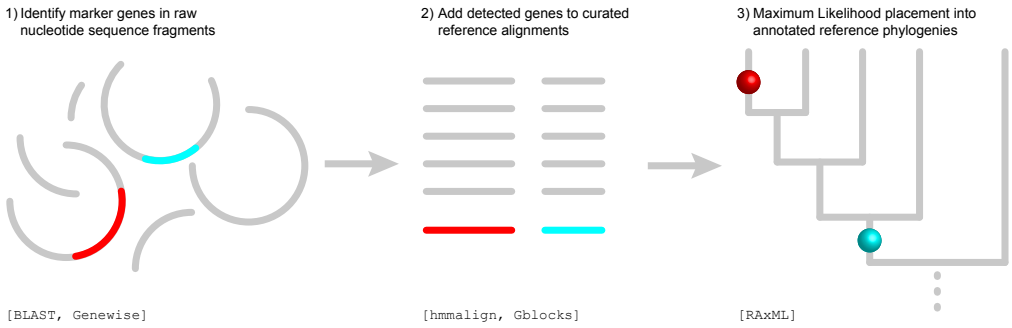


Figure 4.1: The MLTreeMap pipeline. MLTreeMap extracts informative marker genes from the query sequences, aligns them to a set of reference genes and places them into externally provided phylogenies (Stark et al. 2010 [1]).

query sequence at a time. One single placement in the reference tree-of-life required about 2 hours, making the tool unsuitable for analyzing even small metagenomic datasets, because they already contain thousands of sequences. The new version of MLTreeMap is able to process a metagenomic dataset with 20'000 sequences in about 1 hour on our cluster HPC, and in approximately 12 hours on a single CPU. It is available as a web server and as a stand-alone downloadable tool on http://mltreemap.org. We have published it in 2010 [1] (see chapter 9.1).

4.2 MLTreeMap - phylogenetic analysis

4.2.1 Phylogenetic analysis I: protein-coding marker genes

The first phylogenetic analysis of MLTreeMap is based on a set of 40 protein-coding marker genes (table 4.1). They are universal, occur in very low copy numbers per genome (preferably only once) and have undergone very few (or no) horizontal transfers [69]. As can be seen in table 4.1, most of these genes are ribosomal proteins, followed by tRNA synthetases.

COG0012 Predicted GTPase	COG0016 PheS
COG0018 ArgS	COG0048 Ribosomal protein S12
COG0049 Ribosomal protein S7	COG0052 Ribosomal protein S2
COG0080 Ribosomal protein L11	COG0081 Ribosomal protein L1
COG0085 RpoB	COG0087 Ribosomal protein L3
COG0088 Ribosomal protein L4	COG0090 Ribosomal protein L2
COG0091 Ribosomal protein L22	COG0092 Ribosomal protein S3
COG0093 Ribosomal protein L14	COG0094 Ribosomal protein L5
COG0096 Ribosomal protein S8	COG0097 Ribosomal protein L6P/L9E
COG0098 Ribosomal protein S5	COG0099 Ribosomal protein S13
COG0100 Ribosomal protein S11	COG0102 Ribosomal protein L13
COG0103 Ribosomal protein S9	COG0124 HisS
COG0172 SerS	COG0184 Ribosomal protein S15P/S13E
COG0185 Ribosomal protein S19	COG0186 Ribosomal protein S17
COG0197 Ribosomal protein L16/L10E	COG0200 Ribosomal protein L15
COG0201 SecY	COG0202 RpoA
COG0215 CysS	COG0256 Ribosomal protein L18
COG0495 LeuS	COG0522 RpsD
COG0525 ValS	COG0533 QRI7
COG0541 Ffh	COG0552 FtsY

Table 4.1: List of the protein-coding phylogenetic marker genes used by MLTreeMap.

Based on these genes, we have extended the tree-of-life by Ciccarelli et al. [69], increasing the number of taxa included from 191 to 267 (Figure 4.2). We further offer phylogenetic placements in the tree constructed by Wu et al. [128].

4.2.2 Phylogenetic analysis II: 16S & 18S rRNA

The second phylogenetic analysis offered by MLTreeMap is based on 16S and 18S rRNA data. The corresponding tree-of-life phylogeny was constructed using RAxML, applying the settings recommended in chapter 5.1 of the RAxML 7.0.4 manual. As can be seen in Figure 4.3, we were able to retrieve most taxonomic classes described in literature [129, 130]. Nevertheless some of them, such as the firmicutes or the delta proteobacteria, are not monophyletic in

our tree, although their members cluster close together. This seems to be an intrinsic problem of the annotated taxonomy, because other researchers encountered similar problems with these classes [76, 131].

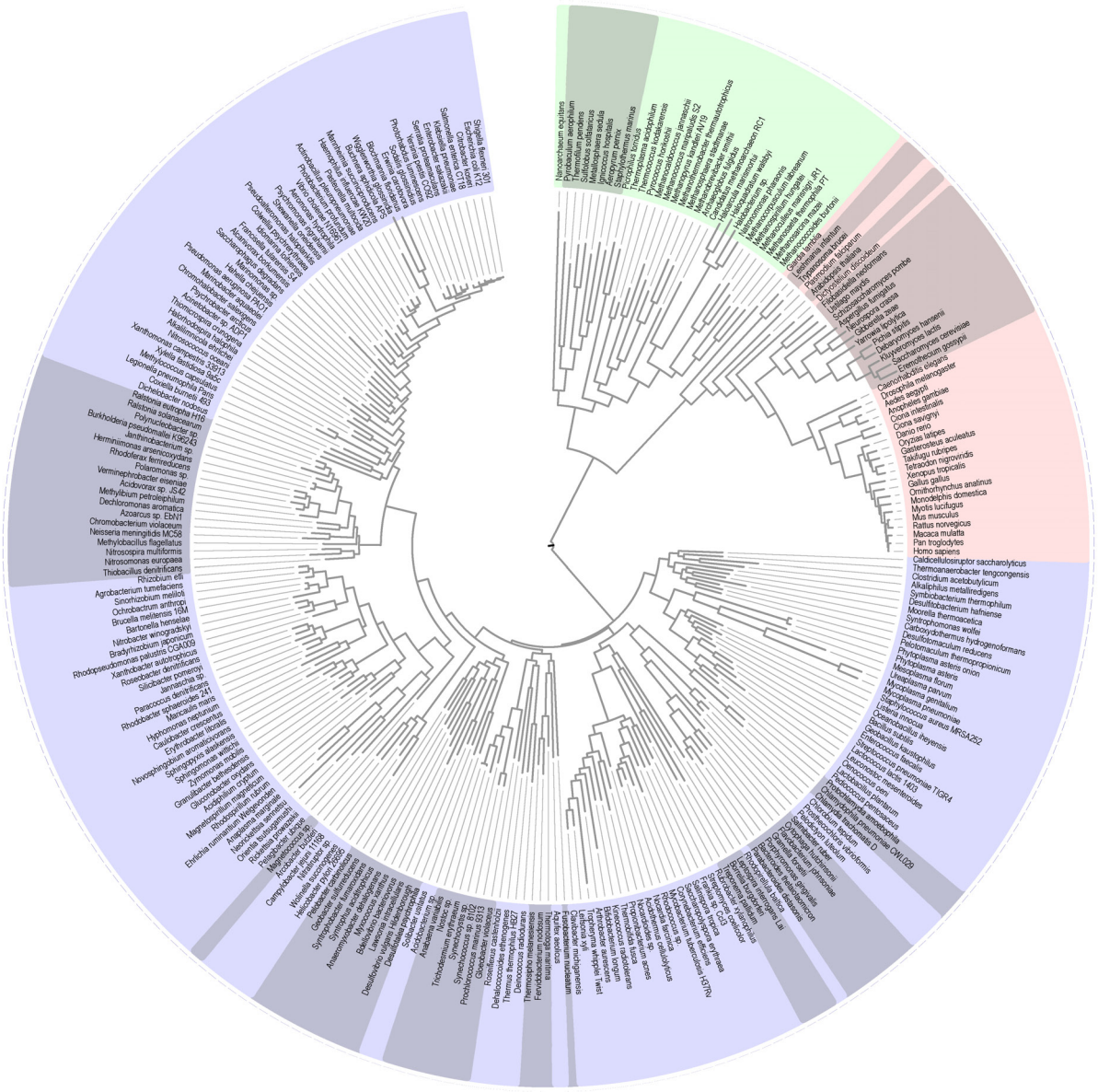


Figure 4.2: Tree-of-life phylogeny I. The tree is based on 40 protein-coding marker genes and contains 267 species (green: Archaea, red: Eukarya, blue: Bacteria).

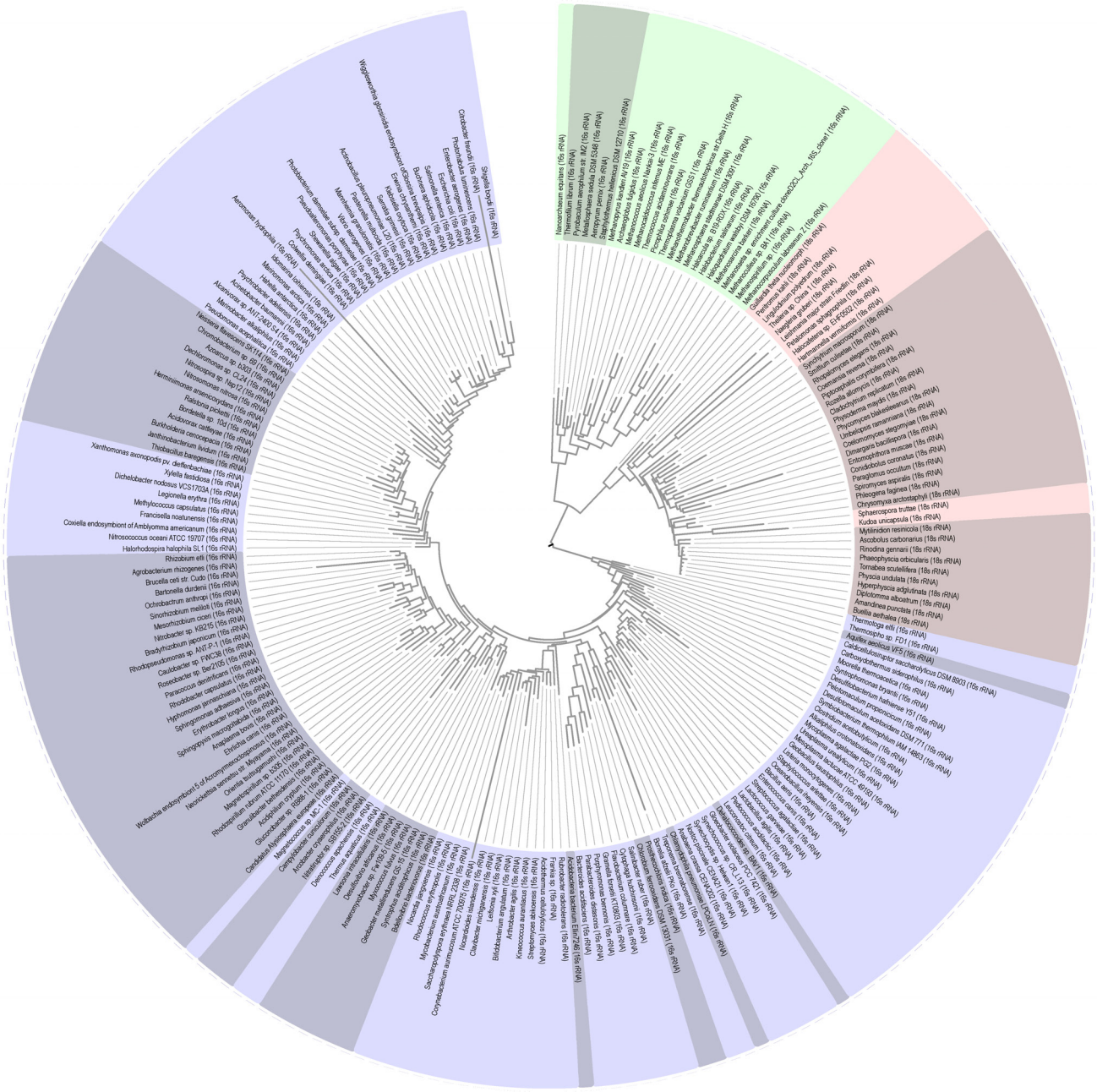


Figure 4.3: Tree-of-life phylogeny II. The tree is based on 16S & 18S rRNA data. It contains 217 species (green: Archaea, red: Eukarya, blue: Bacteria).

4.3 MLTreeMap - functional analysis

In addition to the phylogenetic analysis described above, MLTreeMap also searches for functionally important marker genes. For each of them a gene family phylogeny has been constructed using RAxML (the settings again as described in chapter 5.1 of the RAxML 7.0.4 manual). Afterwards these phylogenies have been compared to literature, in order to annotate insights into functional properties and peculiarities of distinct branches within them.

4.3.1 RuBisCO

RuBisCO is the key enzyme of photosynthesis and essential for the removal of CO_2 from the atmosphere [132]. Due to its low catalytic capacity, photosynthetically active organisms have to produce large amounts of RuBisCO, making it likely to be the most abundant protein on earth [133]. There are several forms of RuBisCO; the most distant one takes no part in photosynthesis and is instead believed to be the evolutionary origin of the protein family [134] (Figure 4.4). The sequences for the reference alignment were obtained from the STRING database [135, 136] (cluster of orthologous genes COG1850).

4.3.2 Nitrogenase

The ability to fix nitrogen from the environment is restricted to a widespread but paraphyletic group of prokaryotes. All of them rely on the enzyme nitrogenase, which is a complex of several subunits (NifH, NifD, NifK, NifE and NifN), for the catalytic process [137, 138]. Even though the nitrogenase gene family has seen various duplications, fusions and horizontal gene transfers, the overall *nif* operon structure is usually highly conserved [137]. Five phylogenetic groups of *nif* genes have been established, of which the first three are functional nitrogenases (note that the labels of group II and III are interchanged in Raymond et al. 2004 [137] and Dekas et al. 2009 [138]. We follow the labeling of the former publication). MLTreeMap contains alignments for the subunits NifH and NifD. The sequences were taken from the KEGG database [139] (orthology groups K02588 and K02586 respectively) (Figures 4.5 and 4.6).

4.3.3 Methane & ammonia monooxygenase

Methane monooxygenase (MMO) is the key enzyme of Methane fixation and thus essential for methanotrophic bacteria [140]. After having constructed a MMO tree based on the KEGG orthology group K08684, we replaced it at the end of 2010 with a tree based on the corresponding sequences obtained from the Functional Gene Pipeline / Repository (FGPR) (available at http://fungene.cme.msu.edu/). This new tree contains sequences of the particulate methane monooxygenase alpha subunit (pmoA) and the ammonia monooxygenase (amoA) (Figure 4.7). The amoA sequences were added because the enzymes are close homologs and both can use ammonium as well as methane as a substrate [141–143]. Thus this phylogenetic tree stands for two very important metabolic pathways: Methane fixation and nitrification (ammonia oxidation).

4.3.4 Reverse dissimilatory sulfite reductase

The reverse dissimilatory sulfite reductase (DsrAB) is one of the key enzymes of the sulfur circle [144]. There are two forms of DsrAB; the first is used by sulfur-oxidizing prokaryotes (SOP) and the second by sulfite-reducing prokaryotes (SRP) for it catalyzes the opposite reaction. Our tree is based on the sequences published by Loy et al. [145]. While we could reproduce the overall tree structure established in their paper, we reject the grouping in 'bacterial' and 'archaeal' DsrAB, due to its internal inconsistency (Figure 4.8).

4.3.5 Cryptochromes and Photolyases

Cryptochromes are photoreceptors, which occur in bacteria and eukaroytes, but have not yet been detected in archaea [146]. They are similar in sequence and structure to the photolyases, which are a group of light activated DNA repair enzymes. Both have been used in the 'subfamily approach' by Singh et al. 2009 [147] to detect functional novelty. The sequences were obtained from the STRING database based on the information given in the aforementioned publication. Of the four named gene families, we only used COG3406 and COG0415. The sequences of COG4338 were too short to serve as suitable markers and NOG16378 does not exist anymore in STRING 8.3.

Figure 4.4: RuBisCO. Forms I - III of RuBisCO show carboxylase activity. Forms I and II belong to the calvin cycle, and form III to the RuPP pathway. Form IV RuBisCO shows no carboxylase activity and its function is not yet entirely clear, but at least some variants are likely to take part in the methionine salvage pathway [134].

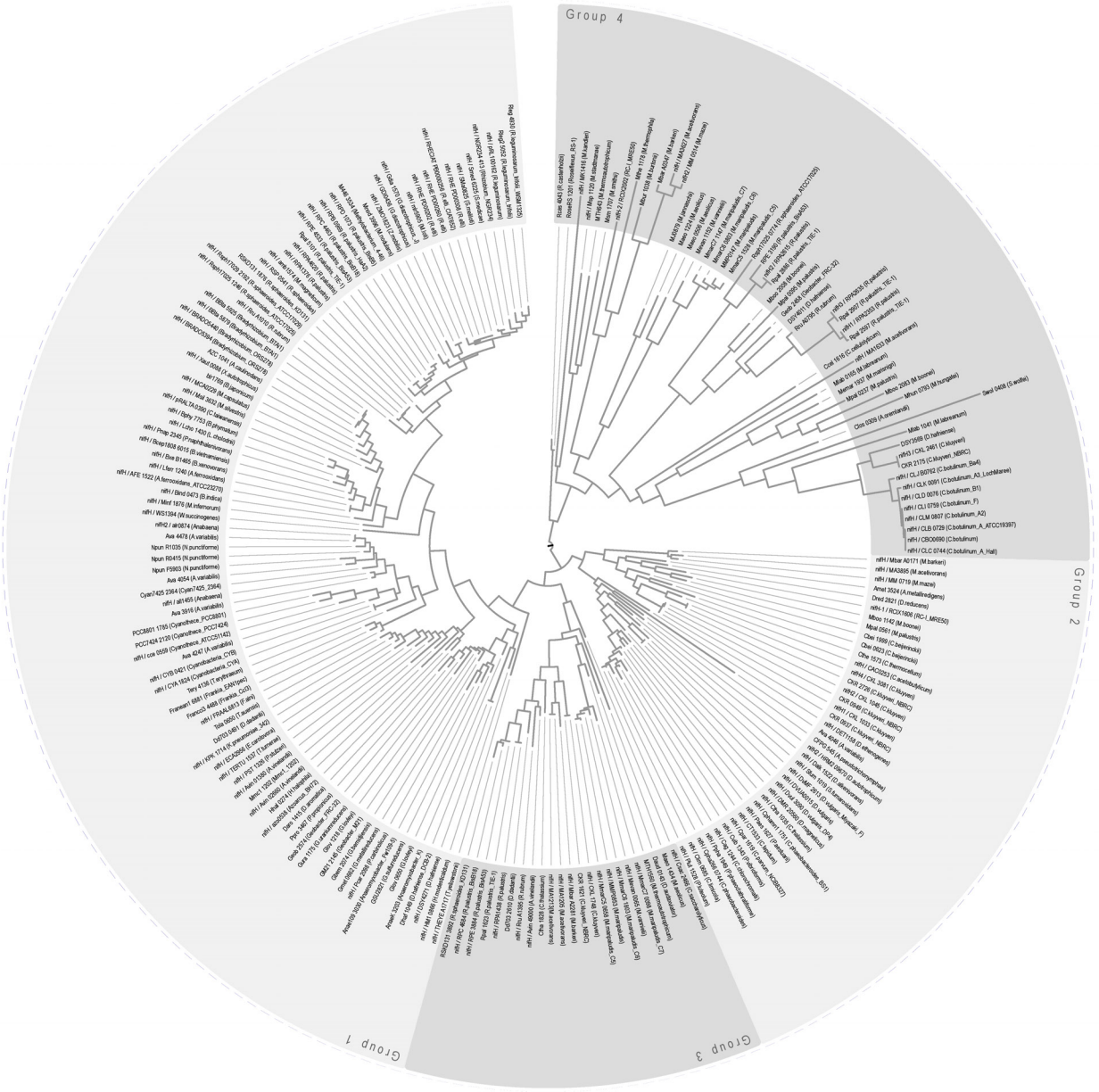


Figure 4.5: NifH. Group I & II nitrogenase proteins are Mo-dependent. Group III is Mo-independent and group IV uncharacterized [137].

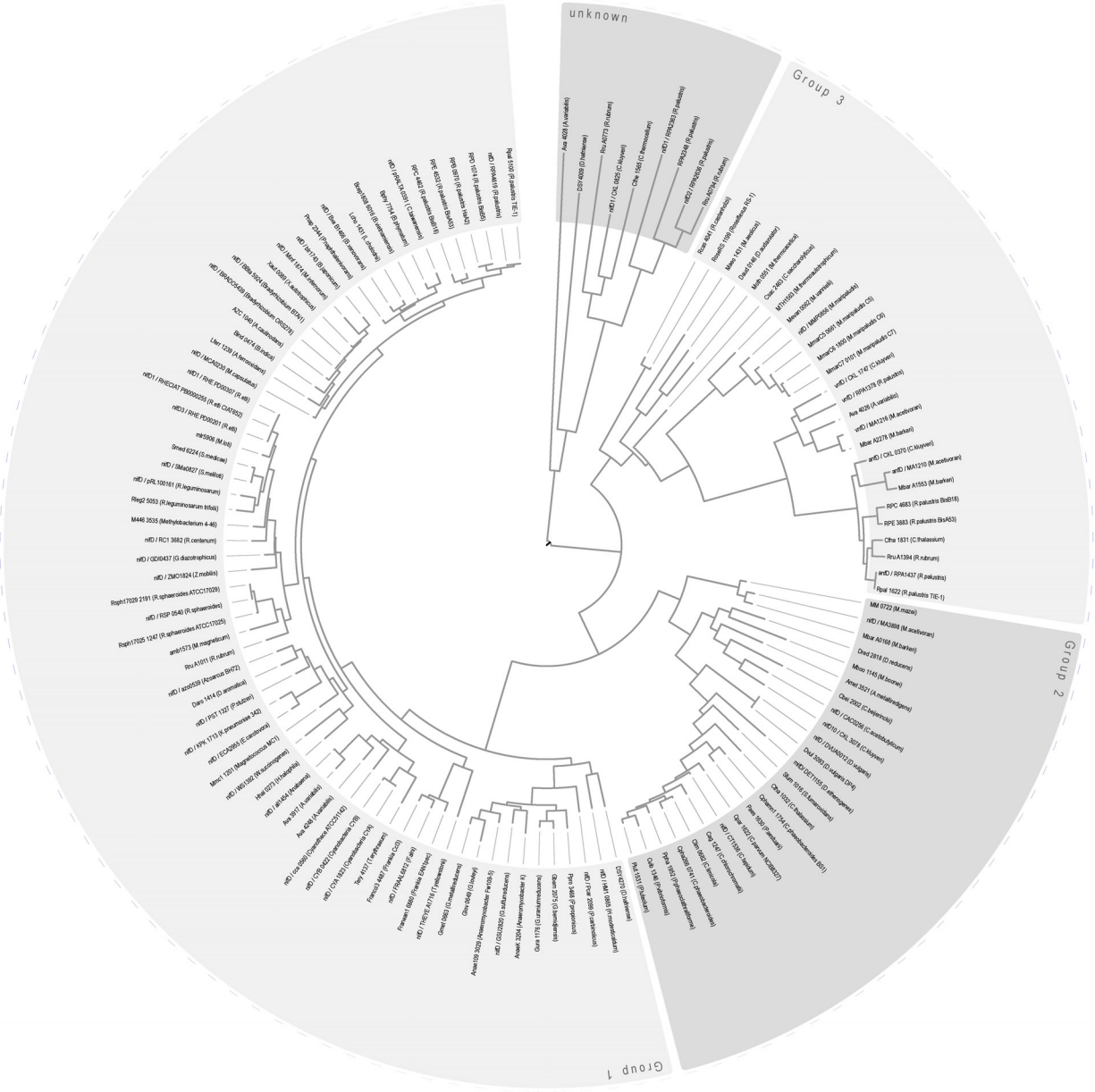


Figure 4.6: NifD. Group I & II nitrogenase proteins are Mo-dependent. Group III is Mo-independent [137].

Figure 4.7: Methane & ammonia monooxygenase. The tree contains sequences of the particulate methane monooxygenase apha subunit (pmoA) and the ammonia monooxygenase (amoA).

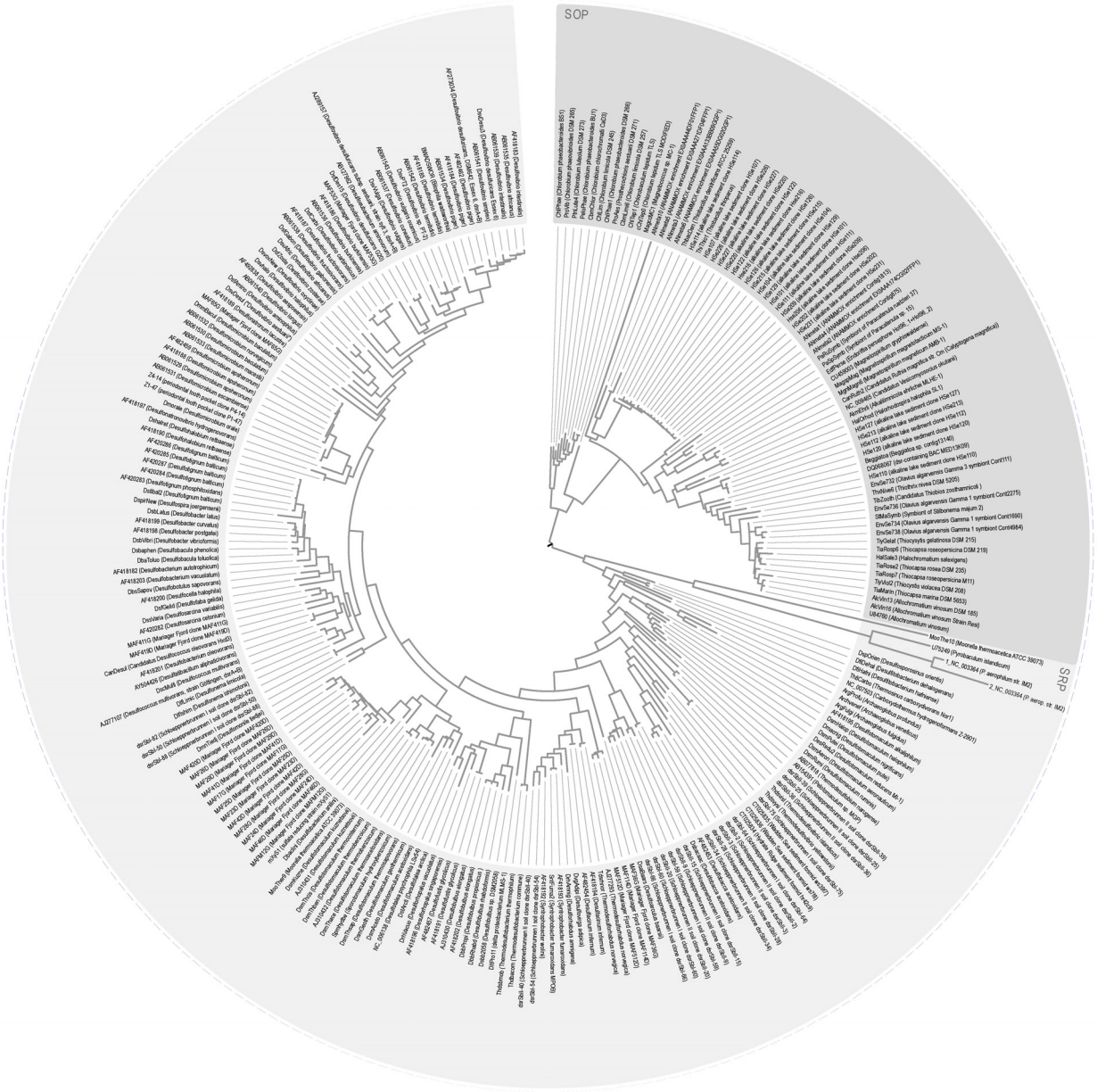


Figure 4.8: Reverse dissimilatory sulfite reductase (DsrAB). The two main groups of DsrAB genes can be clearly distinguished: the first belongs to sulfur-oxidizing prokaryotes (SOP) and the second to sulfate-reducing prokaryotes (SRP).

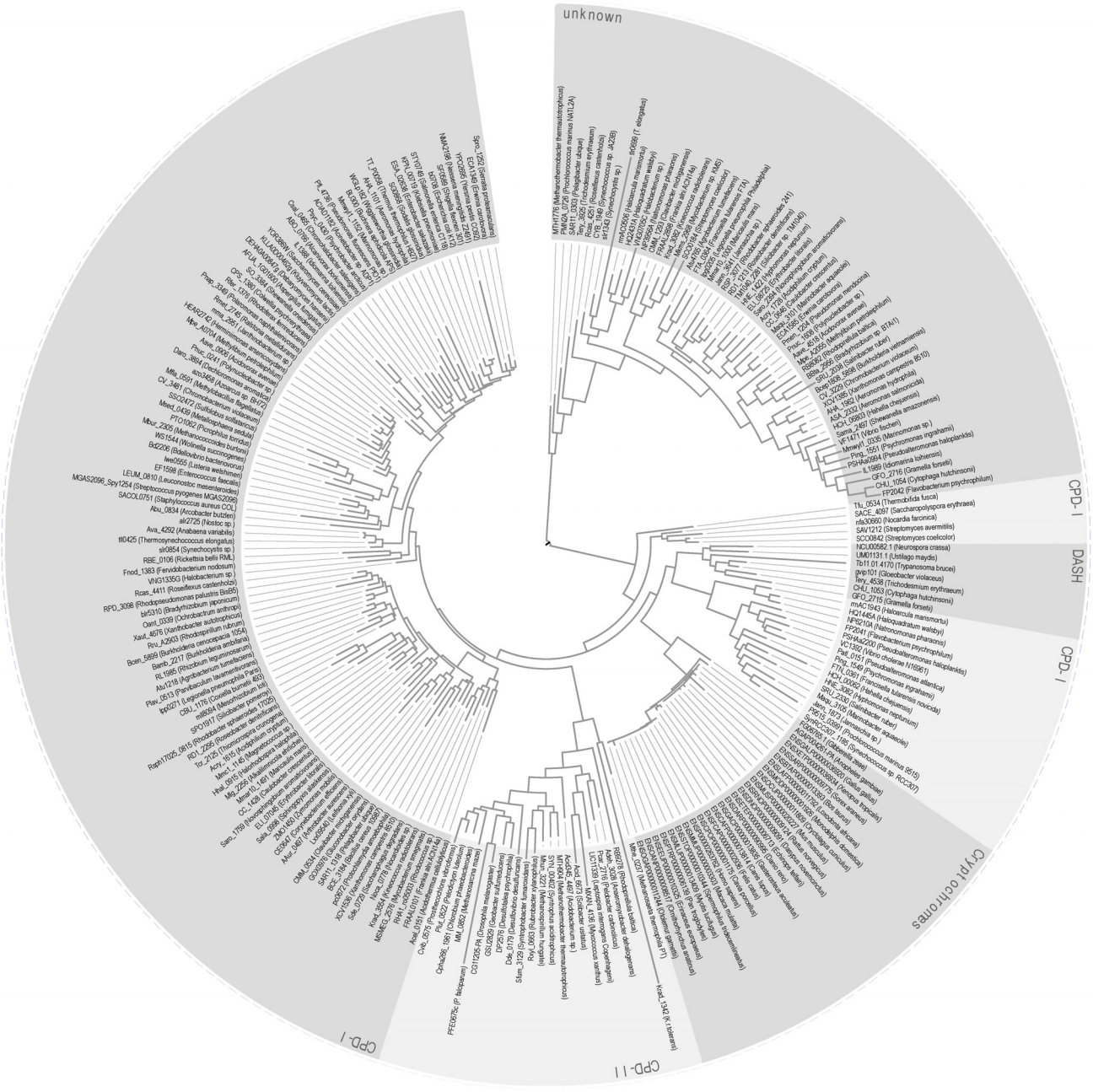


Figure 4.9: Photolyoase & cryptochromes. We could recover the main groups of photolyases (CPD I and CPD II, the first not monophyletically) and cryptochromes (cyryptochromes and DASH). Additionally there is a large homologous group of not yet characterized proteins (see also [147]).

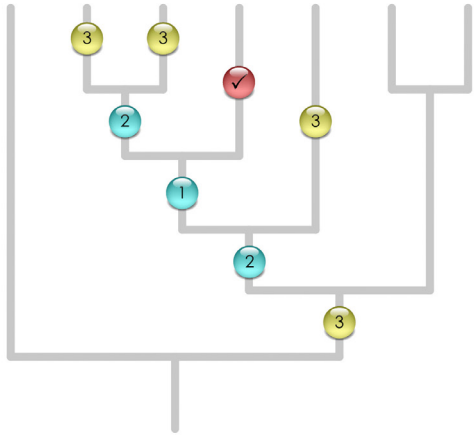
Chapter 5

Validating the MLTreeMap algorithm

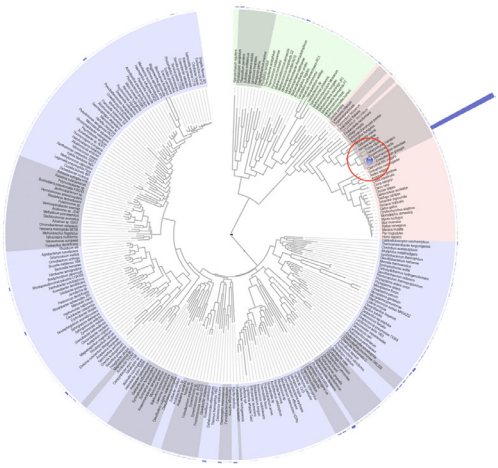
Validation of the MLTreeMap algorithm was an important task during all stages of development. The results of the most thorough testing were included in our publication (see chapter 9.1 for the details). As we have seen in the introduction, most organisms in a metagenomic dataset are not annotated and thus the actual species composition is unknown. This imposes a general problem for all attempts to validate metagenomic algorithms, because their accuracy is not directly measurable. Usually, this problem is solved by creating artificial metagenomes with known phylogenetic origins for all sequences they contain [99]. We started to construct our own artificial datasets by downloading the complete genomes of 85 organisms, which are part of the reference set of MLTreeMap (11 archaea, 64 bacteria and 10 fungi). The genomes were cut into sequences of 1'000 bp each (approximately the length of a Sanger read) and then analyzed with MLTreeMap. As can be seen in Figure 5.1, MLTreeMap produces highly accurate assignments in case of archaeal and bacterial sequences (93-97% of all assignments correctly identified the phylogenetic sequence origin). In case of fungal sequences, the accuracy is lower (57% correct assignments). Nevertheless, after concatenation of the results, the corresponding fungi emerged clearly, because the wrong assignments were scattered all over the tree-of-life phylogeny with very small individual placement weights (e.g. S. cerevisiae in Fig 5.1B).

Mavromatis et al. published three sets of artificial metagenomes, which they proposed as benchmarking tools [99]. The datasets were constructed

A



B



C

Dataset	n	correct	dist 1-2	dist 3	dist 4+
Archaea	503	97.81%	1.19%	0.20%	0.80%
Bacteria	2834	93.79%	4.02%	1.09%	1.10%
Fungi	849	57.60%	1.65%	1.18%	39.57%
simLC contigs	149	84.56%	8.05%	4.70%	2.69%
simMC contigs	210	78.10%	10.95%	9.52%	1.43%
simMC contigs	232	79.74%	9.91%	5.17%	5.18%

Figure 5.1: First MLTreeMap validation. A: Assignments of MLTreeMap were classi ed into four groups according to their node distance (i.e. the unweighted path length) to the phylogenetic sequence origin. The rst group contained correct assignments, which exactly matched the phylogenetic sequence origin (red marble). The second group contained assignments with node distances of 1 and 2 (cyan marbles), and the third group assignments with a node distance of 3 (yellow marbles). All other assignments (i. e. with a node distance ≥ 4) were gathered in the last group. B: Assignments of S. cerevisiae sequences within the tree-of-life phylogeny. The red circle highlights the correct assignments. C: Analysis of the datasets according to the groups introduced in A. The number of sequences, on which marker genes have been found, is shown in column n.

based on randomly selected sequencing reads from isolate genomes (available through the IMG system [148]). Combining these sequencing reads allowed for simulating metagenomes from microbial communities with various complexity levels. The dataset simLC represents low complexity, simMC medium complexity and simHC high complexity communities. Mavromatis et al. reported that with tools like PhyloPythia [107], high quality assignments could only be obtained for the low and medium complexity datasets with sequences longer than 8 kb. For a first analysis, we downloaded and analyzed the contig files of all three datasets, assembled with Phrap [149]. Only results for sequences, which could be mapped to the MLTreeMap reference set at the genus level or lower, were validated (98.7%, 99% and 93% of the simLC, simMC and simHC assignments respectively). The results in Figure 5.1 C show that MLTreeMap performs very well on all three datasets. The lengths of the sequences, on which marker genes were detected, varied considerably, the medians being 1297 bp (simLC), 1420 bp (simMC) and 1174 bp (simHC).

This preliminary and very straightforward analysis demonstrated that the MLTreeMap pipeline produces highly accurate phylogenetic classifications of metagenomic data. However, the constraint that all assessed sequences were derived from organisms, which are part of the MLTreeMap reference set, had two important disadvantages: The first was that this setting is highly artificial. As metagenomes are dominated by unannotated organisms, it is unlikely that real environmental sequences have very close relatives in our reference set. Second, BLAST based approaches are often more accurate than MLTreeMap if annotated sequences are among those that are to be phylotyped (see e.g. in Figure 2 of our publication, shown in chapter 9.1). Thus we designed a new test setting, which is closer to reality, and in which the advantage of MLTreeMap over BLAST based approaches should become eminent: Prior to the analysis of our own artificial metagenomes, we removed the corresponding organisms from the MLTreeMap reference. As for the analysis of the Mavromatis datasets, we now considered both singlet and contig data, which increased the number of analyzed sequences and the underlying species diversity. In case of the simMC dataset, the number of organisms in common with our reference was now below 50%. As can be seen in Figures 2 and 3 of our publication, MLTreeMap handles this more complex task very well and

clearly outperforms MEGAN under these conditions.

Chapter 6

MLTreeMap for users

All resources on MLTreeMap are available at http://mltreemap.org.

6.1 The MLTreeMap web-server

The web-server of MLTreeMap provides a very intuitive user interface. Sequences can either be directly entered into the input box at the center of the page or uploaded as a FASTA file (Figure 6.1). Optionally, each MLTreeMap run can be assigned a user-defined job identifier. After submission, the web-server does some checks on the input (the most important being whether it really contains DNA sequences). If the sequences pass this step, the user will be guided to the 'submitted jobs' page, where the ensuing results are displayed (Figure 6.2). They are listed according to job submission time and labelled by a header in bold print (i.e. the optional job identifier or 'unassigned'). Below each header follow the sequences, on which marker genes have been detected. Clicking on the 'view results' link of a job header leads to a summary page, where the results can be downloaded as a zipped tar repository (Figure 6.3). The corresponding link of a particular sequence leads to a page with detailed information about the hit. While the user interface of the web-server provides most relevant input options of MLTreeMap, it also has some constraints to reduce server load: A job is limited to 50'000 sequences with maximum 100'000 bp each. For larger tasks, the MLTreeMap stand-alone version has been developed.

Figure 6.1: The MLTreeMap web-server I. Screenshot from http://mltreemap.org. The interface contains a navigation bar, an input box for submitting sequences and an information box with references and the version number.

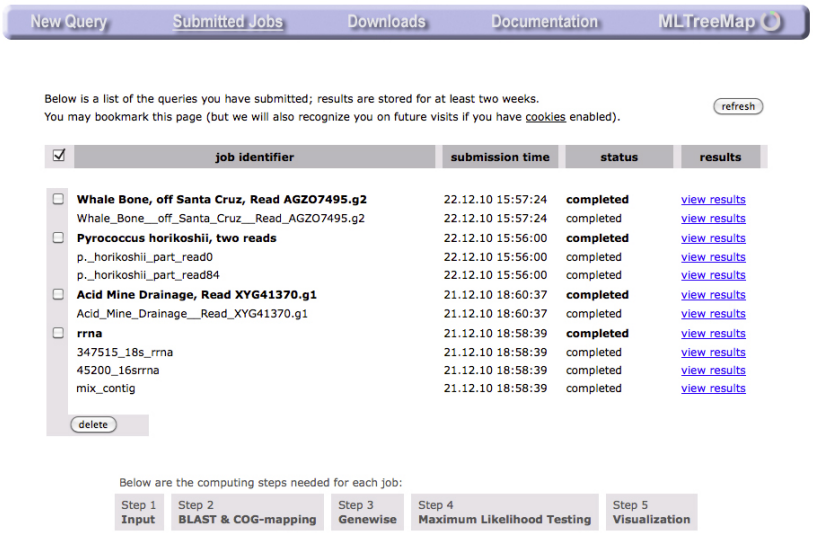


Figure 6.2: The MLTreeMap web-server II. Screenshot from the submitted jobs page of the MLTreeMap web server. The jobs are sorted according to their submission time.

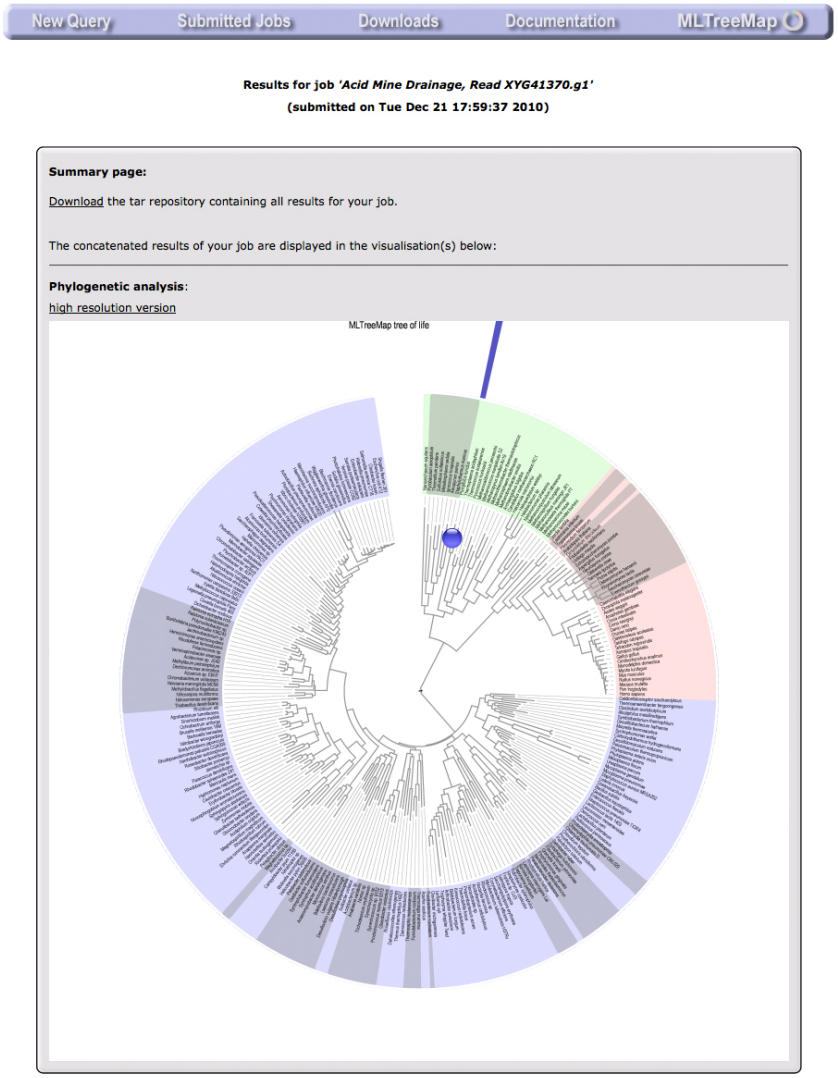


Figure 6.3: The MLTreeMap web-server III. Screenshot of the results summary of a completed MLTreeMap job. A visualization of the concatenated output is displayed below the download link to the tar repository containing all results.

6.2 The MLTreeMap stand-alone version

The first two parts of this section contain information on the downloadable stand-alone modules of MLTreeMap and their motivation. We regard documentations as an essential part of any software and as an important service to the user. Because of this we have written several extensive guides to MLTreeMap: two concern the downloadable modules and a third focuses on modifications of the reference alignments. They are shown in the last three parts of this section. Another more general description of the MLTreeMap algorithm can be found at the documentation page of the web-server, but it will not be reprinted here, since the information that it provides is covered in this thesis.

6.2.1 The MLTreeMap core module

The core module of MLTreeMap is written in Perl with a special focus on minimizing the number of dependencies. It works on Linux and MacOS systems. It is available on the download page of the MLTreeMap web-server together with a detailed documentation (see also chapter 6.2.3). The module already provides a series of pre-installed phylogenetic and functional marker alignments, but it is easy to expand this reference set (see also chapter 6.2.4).

6.2.2 The MLTreeMap imagemaker

In contrast to the MLTreeMap web-server, which visualizes the results automatically, this function is not a part of the stand-alone core module. Instead it is provided by an independent program, called the MLTreeMap imagemaker. This separation was made because of the imagemaker's additional dependencies, mainly resulting from the inclusion of the GD graphics library (available at www.libgd.org). The MLTreeMap imagemaker supports all visualizations known from the web-server, and further provides a mode for displaying different datasets within the same phylogenetic tree, labelled with different colors (see the documentation in chapter 6.2.5 for the details). The module and its documentation are available on the download page of the MLTreeMap web-server.

6.2.3 The MLTreeMap documentation I

A guide to the stand-alone version of MLTreeMap

MLTreeMap is a software framework, designed for phylogenetic and functional analysis of metagenomic data. It searches for instances of marker genes on nucleotide sequences and deduces their most likely origin in a set of reference phylogenies. The current version of MLTreeMap can be downloaded and installed individually. MLTreeMap runs on Mac and Linux systems.
This guide can be downloaded from http://mltreemap.org/treemap_cgi/show_download_page.pl.

A) Installation of MLTreeMap

Step 1

Unzip and unpack the le MLTreeMap_package_2_04.tar.gz .

Step 2

Enter the directory MLTreeMap_package_2_04/install/ .

Type make .

The make program now creates the data structure of MLTreeMap (to be found in MLTreeMap_package_2_04/mltreemap_2_04/) and compiles most needed sub-programs (hmmalign, Genewise and RAxML).

Step 3

BLAST and Gblocks have to be added manually. For this enter the directory MLTreeMap_package_2_04/install/sources/BLAST/ . Here you will nd a collection of BLAST binaries. Choose the one appropriate for your system and copy it to the directory MLTreeMap_package_2_04/mltreemap_2_04/sub_binaries/ .

Repeat this for Gblocks
(to be found in MLTreeMap_package_2_04/install/sources/Gblocks/).

MLTreeMap is now ready to use. You can copy the directory mltreemap_2_04 to any place you like (as well as renaming it). The only dependency of MLTreeMap is that you must have Perl installed on your system.

B) Usage of MLTreeMap

MLTreeMap has to be accessed on the command line.

An example for a valid input command is:
./mltreemap.pl i example_input/rubisco.txt

This will analyze the sequence in the le rubisco.txt and write the result to the output directory mltreemap_2_04/output/ .

Further (optional) input parameters are:

-b number of bootstrap replicates (default: 0 i.e. no bootstrapping).

-c usage a computer cluster (0 = no cluster (default), s = sun grid).

-f RAxML algorithm (v = maximum likelihood (default), p = maximum parsimony).

-g minimal sequence length after Gblocks (default: 50).

-I long input les will be split into les of n sequences each (default: 2000).

-o output directory (default: output/).

-s minimum bitscore for the blast hits (default: 60).

-t phylogenetic reference tree (p = MLTreeMap tree (default), g = GEBA tree).

C) The MLTreeMap output

MLTreeMap searches for phylogenetic and functional marker genes. As soon as the results can be assigned to a speci c marker gene, they are labelled accordingly by the rst character of the output les.

Phylogenetic analysis:

a 16S rRNA reference tree

b 18S rRNA reference tree

g GEBA reference tree

p MLTreeMap reference tree

Functional analysis:

c Photolyase & cryptochrome

d Reverse dissimilatory sul te reductase (DsrAB)

h NifH (K02588)

m Methane & ammonia monooxygenase

n NifD (K02586)

r RuBisCo (COG1850)

Additionally, each output le after the RAxML step gets a header line, providing this information in words. In case of the RuBisCO example from above, the nal output le would look as follows:

Functional analysis, RuBisCO:

Placement weight 100%: Assignment of query to Acry 1067 (Acidiphilium cryptum) (6).

6.2.4 The MLTreeMap documentation II

Adding new reference datasets to MLTreeMap

MLTreeMap is a software framework, designed for phylogenetic and functional analysis of metagenomic data. It searches for instances of marker genes on nucleotide sequences and deduces their most likely origin in a set of reference phylogenies. Part of this set are tree-of-life phylogenies and several functionally important gene families. This guide explains how to add further reference phylogenies to MLTreeMap.
This guide can be downloaded from http://mltreemap.org/treemap_cgi/show_download_page.pl.

To produce sequence placements in reference phylogenies, MLTreeMap needs 4 types of reference les:

- **Alignment files**: these les contain the aligned reference sequences (proteins, not DNA).
- **Hmm files**: a corresponding hmm le belongs to each alignment le.
- **Tree files**: these les contain the reference phylogenies in Newick-tree format.
- **Translation files**: sequence names within the alignment les and the tree les have to be represented by numbers. The translation les allow MLTreeMap to change those numbers back to names.

To create these les and integrate them into MLTreeMap follow the instructions below:

Step 1

Choose a name for your new alignment. It has to be 7 characters long. In the following examples I will use markers as name. You can exchange it with any wording (as long as it is 7 characters long), but if lenames are concerned, the rest should be kept as it is given in the guide.

Choose your marker genes. Create a le called tax_ids_markers.txt. In this le you assign a number to each marker gene name (separate them with a tab).
e.g.
1 marker1
2 marker2
etc.

Now rename your marker genes as follows:
marker1 becomes 1_markers (i.e. its number, followed by an underline and the name of the entire alignment. MLTreeMap will need this information for parsing later on).

Step 2

Align the marker genes in FASTA format (we use MUSCLE to do so). As a result you should get a le somewhat like this:

>1_markers
MATNNVV SELYQLA
>2_markers
MMATTNVV ELYQLA
etc.

Name the le according to the alignment name, which you have chosen in step 1 and ad .fa. In our example the name would be markers.fa.

Format this le using formatdb (available from NCBI). This is necessary for the BLAST step of MLTreeMap.
./formatdb i markers.fa

Step 3

Use hmmbuild (available at http://hmmer.org/) to create the hmm le:

./hmmbuild s markers.hmm markers.fa

Step 4

Use a software (for consistency with the MLTreeMap pipeline preferably RAxML) to create a phylogenetic tree based on your alignment. Save this tree in Newick format. Let us call it markers_tree.txt. It should look somewhat as follows:

((1:0.333, 2:0.323), ...);

NOTE: The tree has to be rooted. If you have an unrooted tree, you can root it manually (e.g. (A,B,C); becomes ((A,B),C);) or by using iTOL (http://itol.embl.de/upload.cgi).
NOTE 2: The sequence names within the tree should be represented by their numbers, which you de ned in step 1 (i.e. marker1 is 1, marker2 is 2 etc.).

Step 5

Copy the les to the following places in the MLTreeMap directory:
cp markers.fa MLTreeMap_home/data/alignment_data/
cp markers.fa MLTreeMap_home/data/geba_alignment_data/
cp markers.hmm MLTreeMap_home/data/hmm_data/
cp markers.hmm MLTreeMap_home/data/geba_hmm_data/
cp markers_tree.txt MLTreeMap_home/data/tree_data/
cp tax_ids_markers.txt MLTreeMap_home/data/tree_data/

Further, an entry is needed in the le MLTreeMap_home/data/tree_data/cog_list.txt. To the last section (#functional_cogs) add a line similar to this (tab delimited):

markers y

This information tells MLTreeMap that there is an additional analysis to be done, based on the markers alignment. The names of the les containing results for this analysis will begin with y_ (e.g. y_concatenated_RAxML_outputs.txt).

NOTE: This guide describes how to add a reference phylogeny, which is based on only one alignment le. If you want to add a reference phylogeny based on several alignment les (similar to our tree-of-life phylogenies), it will be easiest to replace the phylogenetic cogs in cog_list.txt with your new cogs and thus to abolish the traditional tree-of-life analysis of MLTreeMap.

6.2.5 The MLTreeMap documentation III

A guide to the MLTreeMap imagemaker

MLTreeMap is a software framework, designed for phylogenetic and functional analysis of metagenomic data. It searches for instances of marker genes on nucleotide sequences and deduces their most likely origin in a set of reference phylogenies. The MLTreeMap imagemaker has been designed to visualize the placement results within the reference phylogenies. It is not part of the MLTreeMap stand-alone package and has to be installed individually.
This guide can be downloaded from http://mltreemap.org/treemap_cgi/show_download_page.pl.

A) Installation of the MLTreeMap imagemaker

The MLTreeMap imagemaker consists of Perl scripts and has the following dependencies:

GD

Math::Trig

If these dependencies are satis ed, the MLTreeMap imagemaker is ready to use.

B) Usage of the MLTreeMap imagemaker

The MLTreeMap imagemaker has to be accessed on the command line and needs MLTreeMap output les as input.

An example for a valid input command is:
./mltreemap_imagemaker.pl i example_input

This will concatenate all les of the same analysis type into one le and then generate the pictures (they will be written to the output directory
MLTreeMap_imagemaker_2_04/output/).

If you want to enter a single le as input, this is also possible:
./mltreemap_imagemaker.pl i example_input/p_E_coli_RAxML_parsed.txt

Optional input parameters are:

-b parameter for the size of the placement bubbles (default: 0.9).

-d use di erent colors for di erent datasets (0: don t use this mode (default), 1: use this mode). See below for a more detailed description.

-o use this option if you want to write the output to another directory than the default.

-r display 16S and 18S rRNA hits in di erent trees (default: 2) or in one tree-of-life (1). Note: MLTreeMap treats the 16S and 18S rRNA reference data as two di erent trees. But for displaying reasons it might make sense to print all results into a single tree-of-life.

C) Detailed information about the –d input option.

The d option is designed to print the concatenated results from di erent datasets into one single picture, while each dataset is represented by its own color. For the principle see Figure 6.4.

This mode works only if you enter a directory as input e.g. i example_input (if you enter only one le as input, the d option is irrelevant).

Currently the imagemaker supports up to 4 datasets. If your input directory contains more than 4 les of the same type of analysis, the program will die with an error message. You can expand the number of supported datasets by adding more colors in RGB format to the le tree_data/available_dataset_colors.txt.
The input les are alphanumerically assigned to the content of this le.
E.g. Let us assume that the color list contains blue, green and red as follows:
0 0 255
0 255 0
255 0 0
Let us further assume that your input les are the same as in Figure 6.4. If that is the case, then h_ le_1... will get the blue, h_ le_2... the green and h_ le_3... the red color.

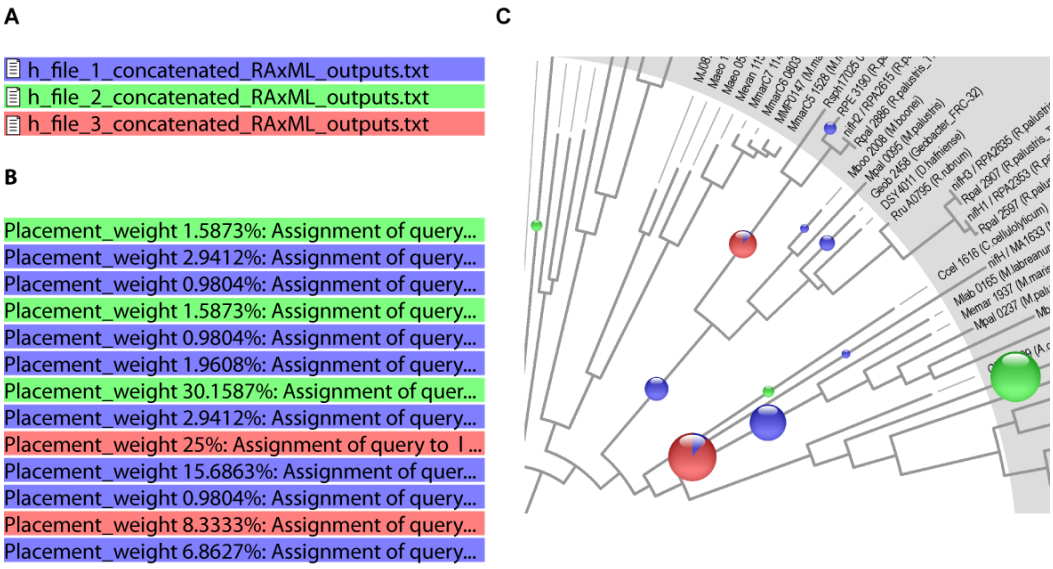


Figure 6.4: The –d input option. Usage: ’./mltreemap_imagemaker.pl –i your_input –d 1’. A: If you use the d option, each le in your_input will be treated as an individual dataset (as indicated by the colored bars). B: Before concatenating the les, each line is labelled with a color code, representing its origin. C: The output picture is drawn according to this color code. If several datasets have placements at the same nodes, pie chart bubbles are drawn to represent the fractional weights.

Chapter 7

Outlook

The last decades have seen a tremendous increase of technical progress in all areas. Metagenomics would not have been possible without the development of high throughput techniques, and it has created the demand for a new generation of computational tools. The task of providing such tools has become even more challenging due to the fact that the advances in sequencing speed and cost effectiveness have far outpaced those in computer technology for over five years now (i.e. sequencing technology moves considerably faster than Moore's law) [150]. A shared requirement for all tools in metagenomics is the ability to handle large quantities of fragmented sequence data in a reasonable amount of time. Several ways of reaching this goal have been developed, with different focuses and standards of accuracy, speed and sequence coverage. We have shown that MLTreeMap belongs to the high accuracy spectrum of tools, at the cost of relatively low speed and coverage (MLTreeMap bases its analyses on approximately only 1% of all sample reads). We consider this a very acceptable trade-off, since most sequences do not contain a reliable phylogenetic signal. Future improvements of MLTreeMap can largely be divided into two groups:

1. **The direct scientific value of MLTreeMap**: Regular updates and extensions of the reference alignments are crucial, if MLTreeMap is to continue to provide the scientific community with valuable and novel insights into microbial communities.

2. **The MLTreeMap algorithm and code**: Updates of the MLTreeMap pipeline have to encompass more than just the reference alignments. The source code has to keep pace with new technical developments and requirements as well. There is considerable potential for further runtime optimization and the recently implemented option for job handling on a cluster HPC can also be refined.

The first task at hand concerns the output format of the MLTreeMap imagemaker and thus belongs to the second group of improvements. Our current visualizations are delivered in the Portable Network Graphics (PNG) format, which means that they are pixel-based. This has several disadvantages, even though the pictures look very appealing: One of them is that locating specific species within the trees is very time-consuming because text searches are impossible. It is further rather difficult to edit the pictures for display on posters or in papers. The ratio between memory consumption and information content is not ideal either: The high resolution images require about 3 MB of disk space, while the same images in a vector-based format would need less than 300 KB. The most important drawback of the MLTreeMap imagemaker is that it cannot display trees containing more than 300 leaves properly with the current settings. As future trees will certainly exceed this limit, this problem has to be solved. The most straightforward approach is to increase the picture size, but if we take the points mentioned above into account, this is not a good way of doing it. Using vector-based graphics on the other hand would solve all these problems at once. Because of that, we plan to switch the output format of the MLTreeMap imagemaker to Scalable Vector Graphics (SVG). The recently published Newick utilities [151] appear to be a very suitable framework for this. They are a bundle of shell programs, capable of drawing tree visualizations in SVG. An important feature is that they include CSS maps for modifying the appearance of the output trees. The options provided by these maps include individual coloring of edges and the possibility of adding 'ornaments' to nodes. We can use the former for labeling specific taxa and the latter for printing our sequence assignments into the pictures. As a consequence the MLTreeMap imagemaker script will be redesigned as a wrapper, which creates the CSS maps and launches the Newick utilities. If their portability is good, we consider to include the re-

vised MLTreeMap imagemaker and its functionalities into the MLTreeMap core module. This would allow us to provide an automated visualization of the results also in the stand-alone version, which would especially benefit the computationally less experienced users.

As we have seen in the chapter on the MLTreeMap documentation, users are encouraged to add their own phylogenies to the MLTreeMap reference set. The task is not very difficult and has been successfully accomplished by scientists from other labs, but it still requires a lot of manual work. To improve this situation we plan to provide a novel 'tree automation pipeline', which is supposed to process and include a given set of sequences into both the MLTreeMap core module and the imagemaker. As a first step, this implies automating the procedure that is described in the second part of the MLTreeMap guide (see chapter 6.2.4). For properly prepared datasets with high sequence identities, this should be relatively easy. The greater challenge is to develop an additional algorithm, which is able to handle low quality datasets. This is difficult because processing them requires searching and correcting for minor annotation mistakes and, even more important, the pruning of sequences that do not align properly. The whole workflow has to be logged and the user must be given the opportunity to influence the decisions made by the algorithm (especially those regarding the pruning steps). The tree automation pipeline further has to root the trees provided by RAxML at a suitable place before they can be used by MLTreeMap. For this we intend to provide two options: user-defined outgroup rooting and automated midpoint rooting [152].

Another potential for improvement concerns the sequence examples, which are available at the MLTreeMap web-server. Our log files indicate that they are frequently used and that this often precedes the analysis of real datasets. Thus we assume that many users check out the MLTreeMap algorithm first, before they apply their data to it. Unfortunately our examples do not show how MLTreeMap performs with whole metagenomes, because they contain only one or two sequences each. The motivation for this was that they have to be sufficient to demonstrate the main features of MLTreeMap without causing much server load. To provide a more realistic example, we intend to precompute several recent metagenomic datasets and make the results

accessible on the web-server.

We have met an increasing interest in MLTreeMap since its publication in 2010. Some user requests were easy to answer, while others led to the development of customized modes and algorithms. In one case, we designed a special version of MLTreeMap: Instead of placing sequences into reference phylogenies, this version constructs trees *de novo* that consist of our marker genes and the query sequences which are mapped to them. In another case, we designed a script to rank MLTreeMap results to taxonomic units, as we did it in Figure 3 of our paper (chapter 9.1). The next reference alignment that is to be added to MLTreeMap, has been suggested by a user of our pipeline: It is the hydrazine oxidoreductase (HZO) gene [153,154]. HZO is an important encyme in anammox bacteria, which do anaerobic ammonium oxidation. We further intend to add the hydroxylamine oxidoreductase (HAO) to the alignment, because it is a member of the same gene family [155]. The HZO/HAO tree will complement the information on functional properties of microbial communities provided by our pmoA/amoA tree, because both HAO and amoA are key enzymes in the nitrification reaction [156].

Recently, we have been approached by the developers of the SmashCommunity software pipeline [157] for a very promising collaboration: They propose to include MLTreeMap into SmashCommunity, a stand-alone tool for the analysis of metagenomic data, which is also able to incorporate external software. Contributing to this excellent software tool will be of great benefit for MLTreeMap as it will increase its public use significantly.

Pursuing these tasks and keeping in touch with the users of MLTreeMap, will ensure that our software pipeline is up to the challenges in metagenomics for many years to come.

Chapter 8

Acknowledgements

First I want to thank Christian von Mering for giving me the opportunity to work on MLTreeMap and for his help and support during all stages of the project. I am very happy that I could do my PhD in his group, the nice atmosphere of which is not least due to his guidance.

Next I want to thank Stefanie Wanka. Without her, life in the lab would not have been half as good. I also wish to thank the other lab members for their friendship and support: Manuel Weiss, Samuel Chaffron, Sabine Schilling, Alexander Roth, Milan Simonovic, Andrea Franceschini, Gabriele Haertinger, João Rodrigues, Mingcong Wang and Sebastian Schmidt. Additional thanks to Alister Smith and Sylvia Davis for some last minute checks on the manuscript.

I am greatly indebted to Alexandros Stamatakis for his contributions to MLTreeMap and for many interesting discussions. Without his input and the customized RAxML algorithms designed especially for us, MLTreeMap as presented in this thesis would not have been possible. Further thanks go to Wolf-Dietrich Hardt and Benjamin Misselwitz for the fruitful collaboration on Salmonella host cell invasion. I would like to thank my other collaborators too, whose input helped me much to improve MLTreeMap: Simon A. Berger, Chris FanLu, Claudia Knief, Chien-Chi Lo, Antoine Page, Young C. Song and Shawn Starkenburg.

I also owe my gratitude to the other members of my thesis committee for very interesting and constructive meetings and their support: Michael Baudis, Jakob Pernthaler and Thomas Wicker.

Special thanks go to the 'usual suspects' for their friendship and sharing the ups and downs of PhD life with me: Regula Jenny, Miriam Meli, Daniel Möckli, Gianluca Ravioli, Giorgio Ravioli, Stefan Roffler and Marc Schneider.

Last but not least I want to express my deepest gratitude to my parents, Brigitte and Paul Stark, for their love and care during my entire life.

Chapter 9

Appendix

9.1 MLTreeMap - accurate Maximum Likelihood placement of environmental DNA sequences into taxonomic and functional reference phylogenies

9.1.1 Preface

We published MLTreeMap in BMC genomics in August 2010 [1]. I did the (re-) implementation of the entire MLTreeMap pipeline and conducted the validation testing. Due to the time gap between submission in April and publication in August, the version of MLTreeMap described in the paper (version 2.011) is not as advanced as the one on which the information given in this thesis is based (version 2.04).

9.1.2 BMC Genomics, 2010

Stark et al. BMC Genomics 2010, 11:461
http://www.biomedcentral.com/1471-2164/11/461

BMC Genomics

METHODOLOGY ARTICLE Open Access

MLTreeMap - accurate Maximum Likelihood placement of environmental DNA sequences into taxonomic and functional reference phylogenies

Manuel Stark[1,2], Simon A Berger[3], Alexandros Stamatakis[3], Christian von Mering[1*]

Abstract

Background: Shotgun sequencing of environmental DNA is an essential technique for characterizing uncultivated microbes *in situ*. However, the taxonomic and functional assignment of the obtained sequence fragments remains a pressing problem.

Results: Existing algorithms are largely optimized for speed and coverage; in contrast, we present here a software framework that focuses on a restricted set of informative gene families, using Maximum Likelihood to assign these with the best possible accuracy. This framework ('MLTreeMap'; http://mltreemap.org/) uses raw nucleotide sequences as input, and includes hand-curated, extensible reference information.

Conclusions: We discuss how we validated our pipeline using complete genomes as well as simulated and actual environmental sequences.

Background

In the field of microbial genomics, successful laboratory cultivation of naturally occurring microbes has become a major bottleneck [1-3]; this limits and biases our understanding of the biochemical capabilities and ecological roles of microbes in their habitats. Since cultivation is a prerequisite for standard genome sequencing approaches, we are still lacking genomic information for many important microbial lineages (including entire phylum-level groups [4,5]). In addition, there is a sequencing backlog even for those strains that have been cultivated successfully; this however is being addressed now by directed sequencing efforts that are underway [6,7]. Nevertheless, the severe biases and the large gaps in the worldwide collection of cultivated isolates make it difficult to fully appreciate evolutionary processes and microbial ecology, or to exploit the large repertoire of microbial genes that might be relevant to medicine and biotechnology. While techniques that analyze single cells, such as multiplexed microfluidics PCR [8] or single-cell genome sequencing [9,10], can provide unequivocal genomic data in the absence of cultivation, these methods are still limited in terms of throughput and usability. Thus, the approach that presently generates the largest amount of unbiased microbial genome sequence data is 'metagenomics' ([11]; also termed 'environmental sequencing').

More than 200 metagenomics projects are currently registered [5] at various stages of completion; these address a wide variety of habitats and microbial lifestyles [12-16]. Typically, in such projects, an environmental sample is processed by lysing cells and indiscriminately isolating genomic DNA; the latter is then fragmented and shotgun-sequenced to a desired depth. However, even when employing the latest next-generation, high-throughput DNA sequencing technologies, the large complexity and genomic heterogeneity of natural microbial communities often preclude *de novo* assembly of complete genomes from the data - instead, a large number of short to medium-sized sequence fragments are obtained. From these, quantitative inferences can already be made regarding genome sizes [17,18], recombination rates [19], and functional repertoires [20,21], among others. However, many of the perhaps more important ecological questions require the assignment of the

* Correspondence: mering@imls.uzh.ch
[1]Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, Switzerland
Full list of author information is available at the end of the article

Stark et al. BMC Genomics 2010, 11:461
http://www.biomedcentral.com/1471-2164/11/461

sequence fragments to the microbial lineage they originate from, a process called 'binning' [12,22].

An increasing number of algorithms have been devised for this task; these can largely be divided into two groups. The first consists of 'unsupervised' approaches [23-27], in which sequences are binned using signature-based algorithms that focus on nucleotide compositional signals (reflected in the relative frequencies of short nucleotide 'words'). These approaches require no external reference information *a priori*; instead, they learn to distinguish the major taxonomic groups from the data itself (although subsequent assignment to known taxonomic entities is often done). In contrast, 'supervised' approaches [28-34] require extensive, annotated, external reference information. For the most part, these approaches interpret the results of large-scale homology searches against sequence databases, sometimes followed by phylogeny reconstruction; the external reference information is usually derived from the available fully sequenced microbial genomes. For both types of approaches, the various implementations differ greatly in their speed, accuracy, coverage, ease of installation and use, and in the interpretation and visualization of the results. Owing to the size and nature of the input data, formal phylogenetics algorithms are relatively rarely used in these pipelines, with three exceptions: Maximum Parsimony in [33], Neighbor Joining in [29], and an approximate Maximum Likelihood approach in [34]. That the Maximum Likelihood approach has not been applied more frequently is somewhat surprising, since it is arguably among the most accurate and best-described techniques in phylogenetics [35-38]. One reason for this is presumably the high computational cost of this approach, which makes it difficult to execute for very large numbers of sequence fragments.

Here, we describe a software framework ("MLTreeMap") that does employ full Maximum Likelihood, and which is specifically designed for metagenomics sequences. We significantly reduced the computational costs through algorithmic improvements, as well as through a focus on a restricted (but user-extensible) set of informative gene families. The aim of the framework is to cover the high-accuracy end of the tool spectrum, with a particular focus on consistency across different sources of input data. To achieve this, the package, a) starts from raw nucleotide sequences to avoid inconsistencies arising from different gene-calling strategies, b) corrects for frame-shifts and other errors on the fly to optimally extract marker genes, c) includes searches against 'off-target' reference sequences to avoid the detection of undesired deep paralogs, d) concatenates marker genes when several of them are observed in a given sequence fragment, and e) offers intuitive visualization features, both via the command-line as well as via the web-server. The framework contains hand-curated reference phylogenies and alignments; in the first full release that we describe here (MLTreeMap version 2.011), these references encompass a total of 44 distinct gene families that have been selected to address both taxonomic as well as functional aspects of microbial assemblages.

Results and Discussion

We have previously outlined [31] and used [39,40] a preliminary version of the MLTreeMap pipeline; however, this initial implementation was not designed for deployment, only focused on phylogenetic information, and was computationally very inefficient (it required up to several hours of CPU time to assign a single nucleotide sequence fragment). We have since achieved a more than 100-fold speed-up, mainly by using more efficient pipeline code, and by switching the employed Maximum Likelihood phylogenetics engine from TREE-PUZZLE [41] to RAxML [42,43]. This switch also enabled us to deploy recent optimizations inside RAxML that were specifically devised for this purpose [Berger et al., submitted; preprint available at http://arxiv.org/abs/0911.2852v1]. The basic work-flow of a fully automated MLTreeMap run proceeds as follows (Figure 1): First, a batch of input sequences (i.e., unannotated nucleotide sequences) are searched for the presence of marker genes, by running BLASTX against a curated collection of reference proteins (including 'off-target' proteins where necessary). In a next step, all detected instances of these marker genes are extracted using GeneWise [44], based on Hidden Markov Models (HMMs) that are provided as part of the MLTreeMap pipeline; this establishes protein-coding open reading frames and exhibits some tolerance to sequencing errors such as frame-shifts or gaps. The query proteins are then aligned to the corresponding reference proteins using hmmalign [45], and the resulting alignments are concatenated in case more than one marker gene is located on a given fragment (this latter step only applies to phylogenetic markers). Next, alignments are subjected to mild gap-removal [46]; and subsequently they are submitted to RAxML. There, the sequences are placed in their most likely position within the corresponding reference phylogeny. Importantly, RAxML is instructed to fully maintain the input topology of the reference phylogeny and to keep it fixed during the computations. Upon launching, RAxML initially optimizes the Maximum Likelihood model parameters and computes all branch-lengths of the reference tree, based on the alignment provided. Next, RAxML will insert (and subsequently remove again) the query sequence(s) one at a time into every possible branch of the reference tree, re-

Stark et al. BMC Genomics 2010, 11:461
http://www.biomedcentral.com/1471-2164/11/461

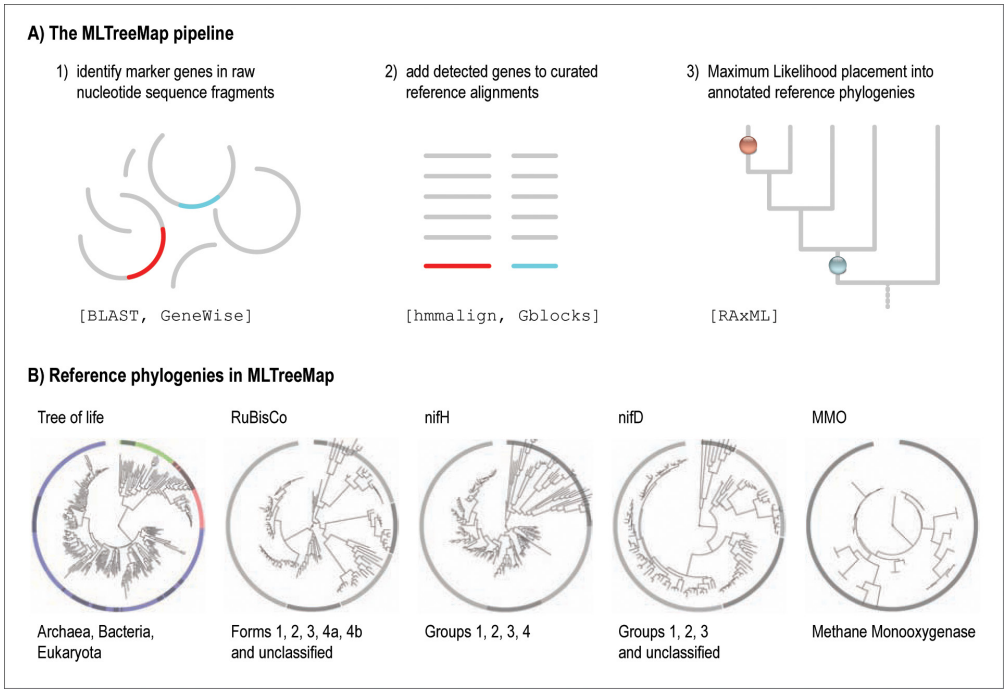


Figure 1 MLTreeMap: Placing anonymous sequence fragments into reference phylogenies. Top: overview of the procedure. Informative marker genes (or fragments thereof) are automatically extracted from raw, un-annotated nucleotide sequence fragments, aligned to reference sequences and then placed into externally provided gene trees using RAxML. Below: Overview of reference phylogenies that are currently available in MLTreeMap.

optimizing the three branch lengths at the insertion position for each attempt. The best-scoring position (branch) for each query sequence is then reported. Optionally, RAxML can use non-parametric bootstrap to account for placement uncertainty. For the bootstrap replicates, heuristics are deployed that only assess the top 10% most promising placement branches as computed on the original (non-bootstrapped) alignment and thereby reduce run times for bootstrap placements by one order of magnitude. Note that, under the settings chosen for MLTreeMap, the actual likelihood computations in RAxML follow the standard Maximum Likelihood approach under a standard protein evolution model, for maximum accuracy. Finally, the results are aggregated, reported in human-readable form and visualized graphically in the context of the reference trees (Figure 1). Currently, 40 of the reference protein families that we provide are collectively used to assess the taxonomic composition of the input sequences (these 40 families were selected based on universal occurrence in all three domains of life, as near-perfect single-copy genes [47]). Another four families serve as indicators for the presence of crucial metabolic pathways (nitrogen fixation, photosynthesis and methane assimilation). In the current implementation, the processing of an amount of DNA sequences that is equivalent to an average microbial genome takes about three to four hours on a single CPU (more when bootstrapping is requested; for example, the above runtime changes to 7 hours when 10 bootstraps are done in each RAxML run). The performance scales roughly linearly with the amount of DNA to be processed; for example, a medium sized metagenome (C1-oxidisers in lake water [48], at 37 Mb) requires about 30 hours to compute on a single CPU; a larger metagenome (220 Mb from a hot spring) requires close to 200 hours. Since the individual DNA fragments can be assessed independently, the pipeline can seamlessly be deployed onto a compute cluster (by splitting the input, and aggregating the results afterwards).

To validate the performance of the MLTreeMap pipeline, we first tested its accuracy on short sequences of known origin. These were generated by artificially

Stark et al. BMC Genomics 2010, 11:461
http://www.biomedcentral.com/1471-2164/11/461

fragmenting fully sequenced genomes into non-overlapping stretches of 1'000 base pairs each (this length corresponds to current read lengths of the Sanger sequencing technology, and it also matches the projected length of the upcoming next release of the 454 pyrosequencing technology). To avoid circularity, we removed the corresponding genomes from our reference alignments and pruned them from the trees. Thus, our testing amounts to leave-one-out cross-validation. Note that our phylogenetic reference tree is already non-redundant at the genus level (with a few exceptions), meaning that removal of the query genome usually results in the next best relative to be available only at the phylogenetic rank of 'family' or higher. The performance of our approach was compared to that of two widely used, previously published approaches, MEGAN [28] and AMPHORA [33], which are based on BLAST searches or Maximum Parsimony insertions, respectively. The algorithmic challenge of our test varies from query genome to query genome, depending on its phylogenetic position (depth) in the reference phylogeny. This is illustrated, for two exemplary genomes, in Figure 2: all three approaches deliver a good accuracy when the query genome remains in the reference (i.e., 95% to 100% of correct placements, see top of Figure 2). However, when removing the query genome from the reference, together with increasingly distant relatives, the accuracy of all three approaches decreases, as expected. This is relevant, because actual environmental sequence fragments will often be fairly unrelated to any fully sequenced genome. Since in our test each query genome is represented by 40 independent reference genes, the resulting placements are spread out over the tree; this is a good visual indication of the nature and extent of the placement error (Figure 2). For the two arbitrary genomes that we chose as examples in Figure 2, Maximum Likelihood and Maximum Parsimony were both performing significantly better than the BLAST-based heuristics implemented in MEGAN. Between the two, Maximum Likelihood performed better in three instances, whereas Parsimony insertion performed better in one instance (note that all pre-processing steps and reference sequences were kept exactly the same for the latter two approaches, in order to facilitate their direct comparison).

We next performed this test systematically, based on 85 complete genomes (11 Archaea, 64 Bacteria and 10 single-celled Eukaryotes (fungi); see Figure 3). This involved testing 406'900 sequence fragments, of which 4'186 were found to contain at least one of our phylogenetic marker genes (i.e., our pipeline typically addresses only about 1% of the sequences in any given sample, by focusing on the most informative parts). We observed that, overall, Maximum Likelihood placed 47.2% of the query sequences at precisely the correct position in the tree, and another 21.3% in close vicinity (i.e., at most two nodes away in the tree). This compares favorably to Maximum Parsimony insertion, using the exact same sequence input (44.8% and 22.0%, respectively). This can also be described in taxonomic terms: Maximum Likelihood places 86.0% of the query sequences within the correct phylum, and 61.2% even within the correct order; these numbers are 83.8% and 55.6% for Maximum Parsimony, respectively. The gain in accuracy over Maximum Parsimony is not dramatic, but it is statistically significant: when re-testing the fragmented bacterial genomes in 1000 bootstrap runs (i.e., randomly sampling genome fragments with replacement), the distributions of accuracy scores for the two approaches were at least four standard deviations apart - testing each of the levels 'phylum', 'order' and 'family'. Overall, there are notable differences with respect to the three kingdoms of life: Bacteria are currently placed with the highest accuracy, with Archaea being a close second, whereas Eukaryotes are assigned with comparatively low accuracy. The difficulties with Eukaryotes can be partly attributed to the presence of more paralogs, and introns (the latter can fragment marker genes), but presumably also to mitochondria and other organelles, which introduce non-eukaryotic versions of the marker genes we employ.

We also assessed our procedure by applying it to entire metagenomics datasets, both simulated [49] and real [50]. For the latter, independent taxonomic information is available, which is based on 16 S ribosomal RNA genes that have been PCR-amplified and sequenced from the very same sample [50]. As is summarized in Figure 4, the results for both datasets are in good quantitative agreement with the known (or measured) composition of the input data. In the case of the simulated dataset [49], the task is necessarily somewhat easier, since this set has been assembled by fragmenting known genomes, and many of these genomes are also contained in our reference phylogeny. Nevertheless, of the 113 genomes that contributed to the 'simMC' dataset [49], more than half (59) are not contained in our reference; and of these, 7 are not even represented at the genus level. In addition, the simulated set contains genomes at widely differing levels of sequence coverage, and the genome sizes are also quite variable (spanning almost one order of magnitude). In spite of this, the overall taxonomic composition is reliably recovered by MLTreeMap, and none of the phyla known to be present in the sample have been missed. For the real metagenomics dataset [50], the actual 'target' composition is not known with much certainty, since the PCR-based assessment that has been reported together with the sample could itself exhibit intrinsic quantitative error.

Stark et al. BMC Genomics 2010, 11:461
http://www.biomedcentral.com/1471-2164/11/461

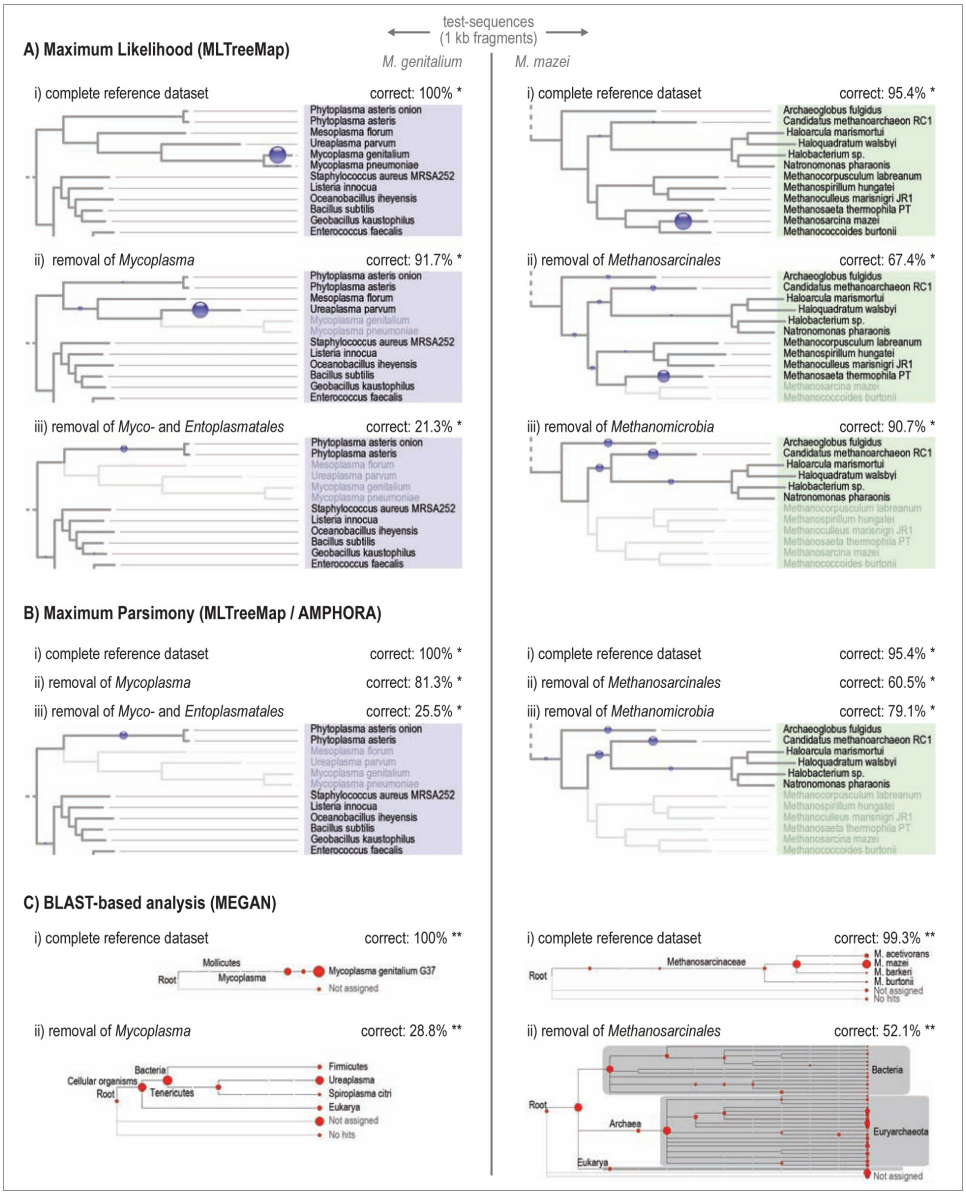


Figure 2 Leave-one-out validation: examples. Individual query genomes were fragmented (1'000 bp fragments) and then placed into reference trees from which the corresponding genomes (or entire clades) had been removed. The assignments are shown graphically (small circles). Note how the placements become increasingly scattered and imprecise upon removal of increasingly deep reference information. MLTreeMap is shown compared to two popular approaches (note that MEGAN, while the least accurate, applies to a much larger fraction of reads in a given sample and thus achieves the best coverage). Definitions of test success: *assignments are designated as correct when they are no more than two nodes away from the target position in the tree. **for MEGAN, assignments are designated as correct when they are mapping to the target phylum.

Stark et al. BMC Genomics 2010, 11:461
http://www.biomedcentral.com/1471-2164/11/461

A) leave-one-out validation
(complete genomes, in 1 kb pieces)

tested domain	nr. of genomes	correct position	distance 1 - 2	distance 3	distance 4 +	correct kingdom	correct phylum	correct class	correct order	correct family
Archaea (ML)	11	60.6%	15.7%	13.3%	10.3%	97.6%	86.7%	78.3%	72.2%	50.1%
Archaea (MP)	11	57.3%	19.9%	13.3%	9.5%	96.6%	85.1%	75.0%	69.4%	47.7%
Bacteria (ML)	64	53.6%	24.3%	6.0%	16.1%	99.8%	93.2%	83.8%	69.4%	53.3%
Bacteria (MP)	64	45.9%	26.7%	7.3%	20.1%	99.6%	90.7%	81.2%	64.5%	47.8%
Fungi (ML)	10	30.2%	23.4%	5.3%	41.1%	67.1%	54.7%	40.0%	36.6%	30.0%
Fungi (MP)	10	25.9%	25.1%	5.2%	43.8%	67.0%	51.4%	36.7%	33.9%	29.2%

assessing inferred tree position (node-distance to actual target)

assessing inferred taxonomy (vs. NCBI classification)

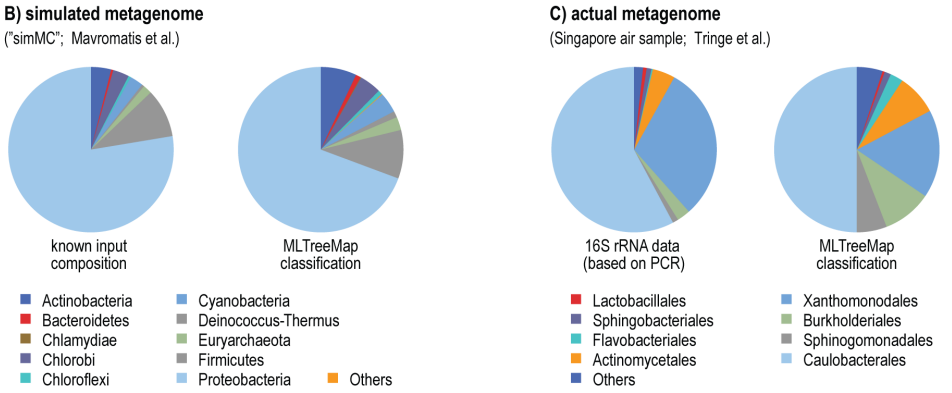


Figure 3 Systematic validation. MLTreeMap is tested on three different types of input (fragmented genomes, as well as simulated and real metagenomes). In all cases, the pipeline has been run with default settings, using the extended reference phylogeny based on Ciccarelli et al. [47].

Indeed, we observe that the MLTreeMap classification appears somewhat more 'balanced' than the PCR-based classification (see Figure 3C: the two most abundant groups make up 88% in the PCR data, but only 67% in the MLTreeMap data). This observation is of course not conclusive: the actual composition of the original sample could well be more biased than reflected in the metagenome. We do note that the distribution of 16 S genes in the metagenome (not PCR-amplified) agrees somewhat better with the MLTreeMap classification than with the PCR-amplified 16 S genes (data not shown), so the observed discrepancy might at least partially be due to the known amplification biases of PCR reactions on mixed templates [51-53], or due to biases in cloning efficiency [54].

Finally, we tested the MLTreeMap pipeline not only with respect to taxonomic assignment, but also with respect to the functional characterization of samples. Currently, the pipeline covers four important enzyme families (RuBisCO, Nitrogenase/NifD, Nitrogenase/NifH, and Methane Monooxygenase). These families are represented by hand-curated alignments, and visualized in the form of annotated protein trees. Future versions of MLTreeMap will extend this set in order to cover a significantly larger number of important diagnostic protein/enzyme families that are indicative of core functions (metabolic and otherwise [55-59]). Figure 4A shows a typical result of MLTreeMap for the functional classification of a set of environmental sequence samples. Three datasets are shown, that each contain representatives of the RuBisCO enzyme family (Ribulose-1,5-bisphosphate carboxylase oxygenase). The mere presence of these genes in the sample could also have been deduced from simple BLAST searches on the data; however, the summary shown in Figure 4A reveals crucial, additional information: first, the mapped sequences

Stark *et al. BMC Genomics* 2010, **11**:461
http://www.biomedcentral.com/1471-2164/11/461

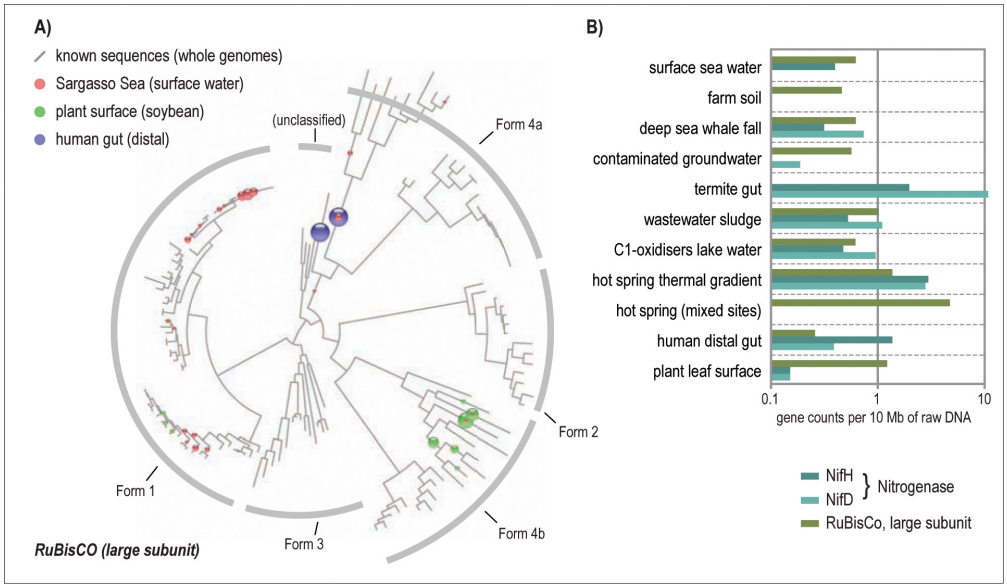


Figure 4 Functional characterization of metagenomes. A) Three published environmental sequence datasets have been searched for instances of the RuBisCo and RuBisCo-like enzyme families, using MLTreeMap. Colored spheres represent sequences mapping to a specific position in the tree, whereby the area of each sphere indicates the relative amount of sequences. The resulting placements are largely non-overlapping, suggesting distinct functional RuBisCo classes encountered/required at each of the environmental sites. B) Several datasets, as available at [69] and [70], were assessed with respect to two metabolic functions (CO^2 fixation, and nitrogen fixation, respectively). All counts were normalized with respect to sampling depth, and are thus directly comparable.

show a clear separation into distinct sub-families of RuBisCO. The surface seawater sample is dominated by subfamily #1, the plant surface sample by subfamily #4b, and the distal human gut by subfamily #4a and other unclassified parts of the tree (subfamilies are designated according to [60]). Second, the functional placements tend to corroborate the taxonomic assignments that MLTreeMaps reports for the same samples (not shown); this enables checks for consistency and/or unexpected horizontal transfers. And third, the placements can be seen to differ dramatically in their distance from the root, that is, in their evolutionary 'depth' with respect to previously known members of the family. For example, in the case of the surface seawater, virtually all sequences were very close to the tips of the tree, in other words closely related to known examples of RuBisCO (mainly from Cyanobacteria and alpha-Proteobacteria). In contrast, instances of RuBisCO-like proteins in the human gut were observed much closer to the root, i.e., at a greater evolutionary distance from previously known sequences and in non-canonical sub-families. From this, it would be much harder to predict their functions, and it is indeed conceivable that they are *not* functioning in CO^2 fixation, but rather in other, possibly sulfur-related metabolic pathways (methionine salvage or yet other, uncharacterized pathways [60-62]). The standardization and ease of use provided by MLTreeMap allow for consistent, semi-quantitative analysis of the functional coding potential of entire collections of metagenomics samples - as an example, Figure 4B shows combined data for 11 distinct metagenomes. In this case, the coding capacities for nitrogen fixation and CO^2 fixation have been compared across samples and sites. Large differences become apparent, including the known paucity of nitrogen fixation genes in some environments [63], but also surprises such as nitrogenase-like genes in the distal human gut. Here again, the availability of the annotated reference trees in the MLTreeMap output is crucial: the sequences are likely of a non-canonical, archaeal type, related to genes in *Methanobrevibacter smithii*, and are thought to function in a process other than nitrogen fixation [64,65].

For both, functional as well as taxonomic assignments, MLTreeMap offers a number of user-definable parameter settings. Users can chose which of two phylogenetic reference trees to use (modified from [7] or [47]),

Stark et al. BMC Genomics 2010, 11:461
http://www.biomedcentral.com/1471-2164/11/461

and whether to use Maximum Likelihood or Maximum Parsimony (the latter works faster but is somewhat less accurate; see Figures 2 and 3). When choosing Maximum Likelihood, users can also request bootstrap replicates. However, bootstrapping will in most cases not be necessary since the input data is already divided into many independent sequence fragments (these constitute 'bootstraps' in some sense; the fragmentation is due to the lack of assembly in most metagenomics projects). Bootstrapping could of course be turned on for specific cases of interest, but for assessing entire datasets it is probably less advisable. This is because individual RAxML runs using all the columns of a given sequence alignment yield more accurate results than each individual bootstrapping run in which columns have been resampled [on average, only 65% of distinct input columns are used in each bootstrap, Berger et al., submitted; this becomes an issue particularly when input sequences are rather short to begin with]. The overall accuracy of MLTreeMap is fairly good already, but it could be further enhanced by improving the coverage and evenness of the reference trees and also by optionally giving deeply assembled contigs (i.e., those with high read coverage) correspondingly more weight in the final aggregation step. Future versions of the pipeline could also likely be optimized further with regards to computational speed - we note that currently much time is still spent outside RAxML, in the pre-processing steps. If further speed-ups can indeed be achieved, then the pipeline should cope well with further advances in sequencing technology - perhaps even to a point in the future when much of the raw data will be discarded immediately after sequencing, and only genes of interest (such as the phylogenetically and functionally informative genes assessed by MLTreeMap) will be kept.

Conclusions

MLTreeMap performs consistent and rapid placements of metagenomics sequence fragments into high-quality, manually curated reference phylogenies - with high accuracy, albeit covering only a restricted fraction of any given sample (around 1%). It focuses on phylogenetically and functionally informative genes, thereby aiming to capture and characterize core aspects of a microbial community. MLTreeMap is one of only a few frameworks that can address microbial eukaryotes on an equal footing with prokaryotes, and it can easily be extended by the user (with any specific gene family of interest). The pipeline will likely be best put to use when analyzing hundreds of samples in comparison: this should ultimately reveal quantitative correlations between certain taxonomic clades and certain functional gene abundance profiles, thus helping to address the classic question of 'who does what' in microbial assemblages.

Materials and methods

Data Sources

Annotated protein-coding genes from fully sequenced genomes were downloaded from STRING [66] and RefSeq [67]. The phylogenetic 'tree-of-life' references were obtained from [7] and [47], but were subsequently modified: we removed genomes for which we were unable to obtain sequences, at the time, and added others. For the tree of [47], we made the representation of organisms non-redundant at the genus level, with a small number of exceptions for fast-evolving genera, and recomputed the best Maximum Likelihood tree, while keeping fixed the original topology of the published tree ('constraints' in RAxML). This computation was based on concatenated alignments of the exact same 40 reference genes as used by MLTreeMap. Note that the purpose of MLTreeMap is not to generate tree-of-life phylogenies *de novo*; instead these trees are provided externally [7,47], we therefore chose to maintain their published topology. For the four functional reference families, gene family information was obtained from KEGG [68] (*nifD: K02586, nifH: K02588, MMO: K08684*) and from STRING [66] (*RuBisCO: COG1850*). In total, the current release 2.01 of MLTreeMap contains 11,069 genes in the reference data; on average, each gene family of interest is represented by 252 genes.

Implementation and Use

MLTreeMap is provided both online (albeit with input-size limitations) as well as offline in form of a command-line executable. The latter is designed with as few external runtime dependencies as possible: BLAST, GeneWise, HMMER and RAxML. Visualization of the results is optional, and a separate Perl-script (with additional dependencies) is provided for this purpose. When using the pipeline, individual reports are generated for each sequence fragment on which marker genes were detected. Aggregated reports are also generated, but this step may have to be repeated by the user (for example when running the pipeline in parallel on separate machines, or when re-weighting the fragments according to additional, external information such as assembly depth or sample size).

The MLTreeMap pipeline has only a few configurable parameters (including: choice of phylogenetic placement method, number of bootstraps, and choice of taxonomic reference phylogeny); other settings are hardcoded with the following default values: required significance of initial BLASTX hits (e = 0.01; database size fixed at 1'000'000), gap removal parameters for Gblocks (-t = p -s = y -u = n

-p = t -b3 = 15 -b4 = 3 -b5 = h -b2 = [0.55 · #alignment_rows]), and required sequence length of the marker genes after alignment and gap removal (50 amino acids). Due to this latter threshold, the pipeline will not yield much useful information for samples with typical read lengths below 300 base pairs (indeed, 500 bp or longer is recommended). The Maximum Likelihood insertion in RAxML is typically done under the following settings: "-f v -m PROTGAMMAWAG" (the WAG substitution model yields the best likelihood scores on the phylogenetic reference trees, compared to all other amino acid substitution models available in RAxML; this was assessed using the RAxML "-f e" option for tree evaluation). For only 7 of the 44 protein families, a substitution model other than WAG is used (RTREV for COG0049, COG0090, COG0092, COG0093 and COG0100; CPREV for COG0201 and BLOSUM62 for Methane Monooxygenase). RAxML works with unrooted trees; however, the MLTreeMap pipeline reports all results in the context of rooted trees, for convenience (the re-rooting is hardcoded for each reference tree). Note that the actual Maximum Likelihood insertion step in MLTreeMap is clearly defined and fairly generic - it could in principle be performed also by software other than RAxML (for example by the PPLACER program; Matsen et al., personal communication; preprint at http://arxiv.org/abs/1003.5943). MLTreeMap can be compiled and executed locally, and previous versions are maintained at our website, for reference (together with the corresponding reference alignments and trees). We plan to update MLTreeMap yearly - each time updating the reference alignments with data from newly sequenced genomes, and extending the repertoire of functional reference families.

Validation

For the validation tests based on whole genomes, the query genomes were artificially fragmented into non-overlapping, consecutive stretches of 1'000 base pairs each. Prior to each test, the respective genome was removed from the reference phylogeny to avoid circularity, and MLTreeMap placements were made using either Maximum Parsimony or Maximum Likelihood (all other settings were identical; bootstrapping was not used). The resulting placements were then compared to the known positions of the query genomes in the reference tree, either by assessing the node distance or the taxonomic assignment. For the latter, the newly placed fragment was assigned to the highest taxonomic rank for which all genomes in the clade below the placement branch were in agreement. For the tests based on simulated metagenomes, we chose the Phrap assembly of the 'medium complexity' simulated dataset, available at http://fames.jgi-psf.org/. The expected target composition of this set is not simply defined by the list of constituent genomes [49]; instead, since the relative genome representation depends on the read coverage of each genome in the simulated set, we weighted all genomes accordingly.

List of Abbreviations

PCR: polymerase chain reaction; ML: Maximum Likelihood; MP: Maximum Parsimony; RuBisCO: Ribulose-1,5-bisphosphate carboxylase oxygenase; MMO: Methane Monooxy-genase.

Authors' Contributions

MS (re-)implemented the entire pipeline, conducted all validation testing and wrote the manuscript. AS developed and implemented the placement algorithms and heuristics in RAxML, defined the interface to the rest of the pipeline, and helped writing the manuscript. SAB supported the systematic validations of the pipeline, validated the improvements in RAxML, and helped writing the manuscript. CVM implemented the initial versions of both the pipeline as well as the website, and wrote the manuscript.

Additional data files

All reference information contained in MLTreeMap (sequences, phylogenies) is available from the associated website http://mltreemap.org/.

Acknowledgements

This work was supported by the Swiss National Science Foundation, by the University of Zurich through its Research Priority Program "Systems Biology and Functional Genomics", and by the Emmy-Noether Program of the German Science Foundation (DFG). We are indebted to Dongying Wu for sharing detailed information regarding his global tree-of-life phylogeny, to Peer Bork for essential scientific and technical support during the early phases of the project, and to Jakob Pernthaler, Thomas Wicker and Michael Baudis for suggestions and critical discussions.

Author details

[1]Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, Switzerland. [2]Ph.D. program in Molecular Life Sciences, University of Zurich and Federal Institute of Technology (ETH), Zurich, Switzerland. [3]The Exelixis Lab, Department of Computer Science, Technische Universität München, Germany.

Received: 23 April 2010 Accepted: 5 August 2010
Published: 5 August 2010

References

1. Alain K, Querellou J: **Cultivating the uncultured: limits, advances and future challenges.** *Extremophiles* 2009, **13(4)**:583-594.
2. Ferrari BC, Winsley T, Gillings M, Binnerup S: **Cultivating previously uncultured soil bacteria using a soil substrate membrane system.** *Nat Protoc* 2008, **3(8)**:1261-1269.
3. Zengler K: **Central role of the cell in microbial ecology.** *Microbiol Mol Biol Rev* 2009, **73(4)**:712-729.
4. Hugenholtz P: **Exploring prokaryotic diversity in the genomic era.** *Genome Biol* 2002, **3(2)**:REVIEWS0003.
5. Liolios K, Chen IM, Mavromatis K, Tavernarakis N, Hugenholtz P, Markowitz VM, Kyrpides NC: **The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata.** *Nucleic Acids Res* 2010, , **38 Database:** D346-354.
6. Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, Bonazzi V, McEwen JE, Wetterstrand KA, Deal C, *et al*: **The NIH Human Microbiome Project.** *Genome Res* 2009, **19(12)**:2317-2323.
7. Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, Kunin V, Goodwin L, Wu M, Tindall BJ, *et al*: **A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea.** *Nature* 2009, **462(7276)**:1056-1060.
8. Ottesen EA, Hong JW, Quake SR, Leadbetter JR: **Microfluidic digital PCR enables multigene analysis of individual environmental bacteria.** *Science* 2006, **314(5804)**:1464-1467.

9. Zhang K, Martiny AC, Reppas NB, Barry KW, Malek J, Chisholm SW, Church GM: **Sequencing genomes from single cells by polymerase cloning.** *Nat Biotechnol* 2006, **24(6)**:680-686.
10. Ishoey T, Woyke T, Stepanauskas R, Novotny M, Lasken RS: **Genomic sequencing of single microbial cells from environmental samples.** *Curr Opin Microbiol* 2008, **11(3)**:198-204.
11. Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM: **Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products.** *Chem Biol* 1998, **5(10)**:R245-249.
12. Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P: **A bioinformatician's guide to metagenomics.** *Microbiol Mol Biol Rev* 2008, **72(4)**:557-578, Table of Contents.
13. Handelsman J: **Metagenomics: application of genomics to uncultured microorganisms.** *Microbiol Mol Biol Rev* 2004, **68(4)**:669-685.
14. Eisen JA: **Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes.** *PLoS Biol* 2007, **5(3)**:e82.
15. Raes J, Foerstner KU, Bork P: **Get the most out of your metagenome: computational analysis of environmental sequence data.** *Curr Opin Microbiol* 2007, **10(5)**:490-498.
16. Tringe SG, Rubin EM: **Metagenomics: DNA sequencing of environmental samples.** *Nat Rev Genet* 2005, **6(11)**:805-814.
17. Raes J, Korbel JO, Lercher MJ, von Mering C, Bork P: **Prediction of effective genome size in metagenomic samples.** *Genome Biol* 2007, **8(1)**:R10.
18. Angly FE, Willner D, Prieto-Davo A, Edwards RA, Schmieder R, Vega-Thurber R, Antonopoulos DA, Barott K, Cottrell MT, Desnues C, *et al*: **The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes.** *PLoS Comput Biol* 2009, **5(12)**: e1000593.
19. Johnson PL, Slatkin M: **Inference of microbial recombination rates from metagenomic data.** *PLoS Genet* 2009, **5(10)**:e1000674.
20. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, *et al*: **Comparative metagenomics of microbial communities.** *Science* 2005, **308(5721)**:554-557.
21. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, *et al*: **A core gut microbiome in obese and lean twins.** *Nature* 2009, **457(7228)**:480-484.
22. McHardy AC, Rigoutsos I: **What's in the mix: phylogenetic classification of metagenome sequence samples.** *Curr Opin Microbiol* 2007, **10(5)**:499-503.
23. Teeling H, Waldmann J, Lombardot T, Bauer M, Glockner FO: **TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences.** *BMC Bioinformatics* 2004, **5**:163.
24. McHardy AC, Martin HG, Tsirigos A, Hugenholtz P, Rigoutsos I: **Accurate phylogenetic classification of variable-length DNA fragments.** *Nat Methods* 2007, **4(1)**:63-72.
25. Abe T, Sugawara H, Kinouchi M, Kanaya S, Ikemura T: **Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples.** *DNA Res* 2005, **12(5)**:281-290.
26. Brady A, Salzberg SL: **Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models.** *Nat Methods* 2009, **6(9)**:673-676.
27. Dick GJ, Andersson AF, Baker BJ, Simmons SL, Thomas BC, Yelton AP, Banfield JF: **Community-wide analysis of microbial genome sequence signatures.** *Genome Biol* 2009, **10(8)**:R85.
28. Huson DH, Auch AF, Qi J, Schuster SC: **MEGAN analysis of metagenomic data.** *Genome Res* 2007, **17(3)**:377-386.
29. Krause L, Diaz NN, Goesmann A, Kelley S, Nattkemper TW, Rohwer F, Edwards RA, Stoye J: **Phylogenetic classification of short environmental DNA fragments.** *Nucleic Acids Res* 2008, **36(7)**:2230-2239.
30. Monzoorul Haque M, Ghosh TS, Komanduri D, Mande SS: **SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences.** *Bioinformatics* 2009, **25(14)**:1722-1730.
31. von Mering C, Hugenholtz P, Raes J, Tringe SG, Doerks T, Jensen LJ, Ward N, Bork P: **Quantitative phylogenetic assessment of microbial communities in diverse environments.** *Science* 2007, **315(5815)**:1126-1130.
32. Dutilh BE, Snel B, Ettema TJ, Huynen MA: **Signature genes as a phylogenomic tool.** *Mol Biol Evol* 2008, **25(8)**:1659-1667.
33. Wu M, Eisen JA: **A simple, fast, and accurate method of phylogenomic inference.** *Genome Biol* 2008, **9(10)**:R151.
34. Schreiber F, Gumrich P, Daniel R, Meinicke P: **Treephyler: fast taxonomic profiling of metagenomes.** *Bioinformatics* 2010, **26(7)**:960-961.
35. Felsenstein J: **Evolutionary trees from DNA sequences: a maximum likelihood approach.** *J Mol Evol* 1981, **17(6)**:368-376.
36. Felsenstein J: **Inferring phylogenies.** Sunderland, Mass.: Sinauer Assoc 2004.
37. Whelan S, Lio P, Goldman N: **Molecular phylogenetics: state-of-the-art methods for looking into the past.** *Trends Genet* 2001, **17(5)**:262-272.
38. Holder M, Lewis PO: **Phylogeny estimation: traditional and Bayesian approaches.** *Nat Rev Genet* 2003, **4(4)**:275-284.
39. Delmotte N, Knief C, Chaffron S, Innerebner G, Roschitzki B, Schlapbach R, von Mering C, Vorholt JA: **Community proteogenomics reveals insights into the physiology of phyllosphere bacteria.** *Proc Natl Acad Sci USA* 2009, **106(38)**:16428-16433.
40. Kunin V, Raes J, Harris JK, Spear JR, Walker JJ, Ivanova N, von Mering C, Bebout BM, Pace NR, Bork P, *et al*: **Millimeter-scale genetic gradients and community-level molecular convergence in a hypersaline microbial mat.** *Mol Syst Biol* 2008, **4**:198.
41. Schmidt HA, Strimmer K, Vingron M, von Haeseler A: **TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing.** *Bioinformatics* 2002, **18(3)**:502-504.
42. Stamatakis A: **RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models.** *Bioinformatics* 2006, **22(21)**:2688-2690.
43. Stamatakis A, Hoover P, Rougemont J: **A rapid bootstrap algorithm for the RAxML Web servers.** *Syst Biol* 2008, **57(5)**:758-771.
44. Birney E, Clamp M, Durbin R: **GeneWise and Genomewise.** *Genome Res* 2004, **14(5)**:988-995.
45. Durbin R: **Biological sequence analysis: probabilistic models of proteins and nucleic acids.** Cambridge [u.a.]: Cambridge Univ. Press 1998.
46. Talavera G, Castresana J: **Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments.** *Syst Biol* 2007, **56(4)**:564-577.
47. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P: **Toward automatic reconstruction of a highly resolved tree of life.** *Science* 2006, **311(5765)**:1283-1287.
48. Kalyuzhnaya MG, Lapidus A, Ivanova N, Copeland AC, McHardy AC, Szeto E, Salamov A, Grigoriev IV, Suciu D, Levine SR, *et al*: **High-resolution metagenomics targets specific functional types in complex microbial communities.** *Nat Biotechnol* 2008, **26(9)**:1029-1034.
49. Mavromatis K, Ivanova N, Barry K, Shapiro H, Goltsman E, McHardy AC, Rigoutsos I, Salamov A, Korzeniewski F, Land M, *et al*: **Use of simulated data sets to evaluate the fidelity of metagenomic processing methods.** *Nat Methods* 2007, **4(6)**:495-500.
50. Tringe SG, Zhang T, Liu X, Yu Y, Lee WH, Yap J, Yao F, Suan ST, Ing SK, Haynes M, *et al*: **The airborne metagenome in an indoor urban environment.** *PLoS One* 2008, **3(4)**:e1862.
51. Baker GC, Smith JJ, Cowan DA: **Review and re-analysis of domain-specific 16 S primers.** *J Microbiol Methods* 2003, **55(3)**:541-555.
52. Polz MF, Cavanaugh CM: **Bias in template-to-product ratios in multitemplate PCR.** *Appl Environ Microbiol* 1998, **64(10)**:3724-3730.
53. Sipos R, Szekely AJ, Palatinszky M, Revesz S, Marialigeti K, Nikolausz M: **Effect of primer mismatch, annealing temperature and PCR cycle number on 16 S rRNA gene-targetting bacterial community analysis.** *FEMS Microbiol Ecol* 2007, **60(2)**:341-350.
54. DeSantis TZ, Brodie EL, Moberg JP, Zubieta IX, Piceno YM, Andersen GL: **High-density universal 16 S rRNA microarray analysis reveals broader diversity than typical clone library when sampling the environment.** *Microb Ecol* 2007, **53(3)**:371-383.
55. Wagner M, Loy A, Klein M, Lee N, Ramsing NB, Stahl DA, Friedrich MW: **Functional marker genes for identification of sulfate-reducing prokaryotes.** *Methods Enzymol* 2005, **397**:469-489.
56. Junier P, Molina V, Dorador C, Hadas O, Kim OS, Junier T, Witzel JP, Imhoff JF: **Phylogenetic and functional marker genes to study ammonia-oxidizing microorganisms (AOM) in the environment.** *Appl Microbiol Biotechnol* 2010, **85(3)**:425-440.
57. Braker G, Zhou J, Wu L, Devol AH, Tiedje JM: **Nitrite reductase genes (nirK and nirS) as functional markers to investigate diversity of denitrifying bacteria in pacific northwest marine sediment communities.** *Appl Environ Microbiol* 2000, **66(5)**:2096-2104.

58. Auguet JC, Borrego CM, Baneras L, Casamayor EO: **Fingerprinting the genetic diversity of the biotin carboxylase gene (accC) in aquatic ecosystems as a potential marker for studies of carbon dioxide assimilation in the dark.** *Environ Microbiol* 2008, **10(10)**:2527-2536.
59. Foerstner KU, Doerks T, Creevey CJ, Doerks A, Bork P: **A computational screen for type I polyketide synthases in metagenomics shotgun data.** *PLoS One* 2008, **3(10)**:e3515.
60. Ashida H, Danchin A, Yokota A: **Was photosynthetic RuBisCO recruited by acquisitive evolution from RuBisCO-like proteins involved in sulfur metabolism?** *Res Microbiol* 2005, **156(5-6)**:611-618.
61. Ashida H, Saito Y, Kojima C, Kobayashi K, Ogasawara N, Yokota A: **A functional link between RuBisCO-like protein of Bacillus and photosynthetic RuBisCO.** *Science* 2003, **302(5643)**:286-290.
62. Imker HJ, Singh J, Warlick BP, Tabita FR, Gerlt JA: **Mechanistic diversity in the RuBisCO superfamily: a novel isomerization reaction catalyzed by the RuBisCO-like protein from Rhodospirillum rubrum.** *Biochemistry* 2008, **47(43)**:11171-11173.
63. Johnston AW, Li Y, Ogilvie L: **Metagenomic marine nitrogen fixation–feast or famine?** *Trends Microbiol* 2005, **13(9)**:416-420.
64. Raymond J, Siefert JL, Staples CR, Blankenship RE: **The natural history of nitrogen fixation.** *Mol Biol Evol* 2004, **21(3)**:541-554.
65. Ohkuma M, Noda S, Kudo T: **Phylogenetic diversity of nitrogen fixation genes in the symbiotic microbial community in the gut of diverse termites.** *Appl Environ Microbiol* 1999, **65(11)**:4926-4934.
66. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, *et al*: **STRING 8–a global view on proteins and their functional interactions in 630 organisms.** *Nucleic Acids Res* 2009, , **37 Database:** D412-416.
67. Pruitt KD, Tatusova T, Klimke W, Maglott DR: **NCBI Reference Sequences: current status, policy and new initiatives.** *Nucleic Acids Res* 2009, , **37 Database:** D32-36.
68. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, *et al*: **KEGG for linking genomes to life and the environment.** *Nucleic Acids Res* 2008, , **36 Database:** D480-484.
69. Markowitz VM, Ivanova NN, Szeto E, Palaniappan K, Chu K, Dalevi D, Chen IM, Grechkin Y, Dubchak I, Anderson I, *et al*: **IMG/M: a data management and analysis system for metagenomes.** *Nucleic Acids Res* 2008, , **36 Database:** D534-538.
70. Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M: **CAMERA: a community resource for metagenomics.** *PLoS Biol* 2007, **5(3)**:e75.

doi:10.1186/1471-2164-11-461
Cite this article as: Stark *et al.*: MLTreeMap - accurate Maximum Likelihood placement of environmental DNA sequences into taxonomic and functional reference phylogenies. *BMC Genomics* 2010 **11**:461.

9.2 The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored

9.2.1 Preface

Establishing protein networks is crucial for the understanding of cellular functions. Nevertheless this is an extremely challenging task because there is a multitude of possible functional connections between proteins and detecting those interactions is not easy. STRING is a database dedicated to presenting integrated information on annotated protein-protein interactions, as well as to predicting them *de novo*. Even though I was mainly involved in the STRING development during my master's thesis [135], I kept supporting the STRING-team during my work as a PhD student. For the 2011 paper I designed a software module, which checked the 1'100 proteomes that were to be used in STRING for the presence of our 40 phylogenetically relevant protein-coding marker genes. As they are ubiquitous in most known organisms, their presence is also an indicator of the quality of the proteomes. Using this pipeline, we detected one proteome, which we had to reject due to bad annotation (Giardia intestinalis). There were two others that possessed only about 20 markers, but in these cases this was likely to reflect biological reality. The reason for this was that these proteomes belonged to organisms, which are obligate intracellular endosymbionts (Candidatus Carsonella ruddii PV and Candidatus Hodgkinia cicadicola Dsem). In addition to this quality check I also gave support to the people working on the confidence scoring system of STRING.

9.2.2 Nucleic Acids Research, 2011

Published online 2 November 2010 *Nucleic Acids Research, 2011, Vol. 39, Database issue* ***D561–D568***
doi:10.1093/nar/gkq973

The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored

Damian Szklarczyk[1], Andrea Franceschini[2], Michael Kuhn[3], Milan Simonovic[2], Alexander Roth[2], Pablo Minguez[4], Tobias Doerks[4], Manuel Stark[2], Jean Muller[5,6], Peer Bork[4,7,*], Lars J. Jensen[1,*] and Christian von Mering[2,*]

[1]Faculty of Health Sciences, Novo Nordisk Foundation Centre for Protein Research, University of Copenhagen, Denmark, [2]Faculty of Science, Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, Switzerland, [3]Biotechnology Center, Technical University Dresden, [4]Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany, [5]Institute of Genetics and Molecular and Cellular Biology, CNRS, INSERM, University of Strasbourg, [6]Genetic Diagnostics Laboratory, CHU Strasbourg Nouvel Hôpital Civil, Strasbourg, France and [7]Max-Delbrück-Centre for Molecular Medicine, Berlin, Germany

Received September 7, 2010; Accepted October 3, 2010

ABSTRACT

An essential prerequisite for any systems-level understanding of cellular functions is to correctly uncover and annotate all functional interactions among proteins in the cell. Toward this goal, remarkable progress has been made in recent years, both in terms of experimental measurements and computational prediction techniques. However, public efforts to collect and present protein interaction information have struggled to keep up with the pace of interaction discovery, partly because protein–protein interaction information can be error-prone and require considerable effort to annotate. Here, we present an update on the online database resource Search Tool for the Retrieval of Interacting Genes (STRING); it provides uniquely comprehensive coverage and ease of access to both experimental as well as predicted interaction information. Interactions in STRING are provided with a confidence score, and accessory information such as protein domains and 3D structures is made available, all within a stable and consistent identifier space. New features in STRING include an interactive network viewer that can cluster networks on demand, updated on-screen previews of structural information including homology models, extensive data updates and strongly improved connectivity and integration with third-party resources. Version 9.0 of STRING covers more than 1100 completely sequenced organisms; the resource can be reached at http://string-db.org.

INTRODUCTION

Proteins can form a variety of functional connections with each other, including stable complexes, metabolic pathways and a bewildering array of direct and indirect regulatory interactions. These connections can be conceptualized as networks and the size and complex organization of these networks present a unique opportunity to view a given genome as something more than just a static collection of distinct genetic functions. Indeed, the 'network view' on a genome is increasingly being taken in many areas of applied biology: protein networks are used to increase the statistical power in human genetics (1,2), to aid in drug discovery (3,4), to close gaps in metabolic enzyme knowledge (5,6) and to predict phenotypes and gene functions (7,8), to name just a few examples.

While clearly very useful, the annotation and storage of protein–protein associations in databases is less straightforward than for other types of data (such as genomic

*To whom correspondence should be addressed. Tel: +49 6221 387 8526; Fax: +49 6221 387 8517; Email: bork@embl.de
Correspondence may also be addressed to Lars J. Jensen. Tel: +45 35 32 50 25; Fax: +45 35 32 50 01; Email: lars.juhl.jensen@cpr.ku.dk
Correspondence may also be addressed to Christian von Mering. Tel: +41 44 6353147; Fax: +41 44 6356864; Email: mering@imls.uzh.ch

The authors wish it to be known that, in their opinion, the first three authors should also be regarded as joint First Authors.

D562 Nucleic Acids Research, 2011, Vol. 39, Database issue

sequence data or taxonomy information). This is because functional interactions between proteins can span a wide spectrum of mechanisms and specificities, often have high error rates and may depend on biological context (such as environmental condition or tissue type). Consequently, considerable information is needed to describe the various aspects of a given protein–protein association and a number of standards have been developed for this purpose with distinct levels of expressivity and specialization (9–13). Likewise, the actual annotations and interaction records themselves are scattered over a number of public resources. Experimental data on physical protein–protein interactions are mostly stored in a group of dedicated databases that together form the International Molecular Exchange (IMEx) consortium (14–21). Annotated pathway knowledge is mostly kept in a separate set of resources (22–24) and yet other interactions can be found in various organism-specific databases (25,26) or text-mining resources (27,28). Furthermore, a number of algorithms have been devised that allow *de novo* prediction of functional links between proteins (29–32), albeit usually with considerable rates of false positives and without providing hints on the specificity and type of a predicted interaction.

Given all these distinct types and sources of protein–protein association information, it is highly desirable for users to have an integration and re-appraisal that can be easily searched and browsed, at one single site. The Search Tool for the Retrieval of Interacting Genes (STRING) database resource aims to provide this service, by acting as a 'one-stop shop' for all information on functional links between proteins. It is by no means the only such site: related resources that are currently being actively maintained include VisANT (33), GeneMANIA (34), N-Browse (35), I^2D (36), APID (37), bioPIXIE (38) and ConsensusPathDB (39). Each of these sites has unique features and distinct strengths and users should carefully compare them for any specific task at hand. The main strengths of STRING lie in its unique comprehensiveness, its confidence scoring and its interactive and intuitive user interface. STRING is the only site to cover hundreds (and soon more than 1100) organisms—ranging from Bacteria and Archaea to humans. This large number of organisms, represented by their fully sequenced genomes, also enables STRING to periodically execute interaction prediction algorithms that depend on exhaustive genome sequence information. The resource also transfers interaction information between organisms where applicable, thereby significantly increasing coverage particularly for poorly studied organisms. The confidence scoring is another key feature of STRING, giving guidance to users who want to balance different levels of coverage and accuracy. Lastly, the unique and compact user interface enables fast and *ad hoc* use of the resource, with a quick learning curve and no need for setup or installation.

Here, we briefly describe the content and procedures currently used in STRING and describe new features that have been added since our last update on the resource (40).

User experience and content

Users enter STRING via its web portal (http://string-db.org) and identify one or more proteins of interest. Various types of identifiers are recognized by the system and a full-text search on gene annotations is conducted in parallel to aid in the identification. Using the search results, STRING will either recognize automatically or ask the user to disambiguate, the organism of interest. The user is then presented with the input protein(s) in the context of a graphical network of interaction partners (Figure 1). From this network, pop-up windows lead to detailed information on each node (or edge) in the network, providing accessory information on a protein or on the evidence behind a proposed connection. The network display can be modified by adding or removing proteins, changing the required confidence level and by selecting or de-selecting certain evidence types (for example, users might choose to filter out the results of computational predictions).

The interactive network viewer in STRING has been re-designed extensively. It is now based on Adobe's Flash Player (version 10 or better is recommended) and allows users to freely reposition nodes in the network. Optionally, this can be done while running a spring-embedded layout algorithm in real time. Upon switching to the 'advanced' mode of the viewer, users can also apply clustering algorithms to the network (41–43), which is then visually partitioned accordingly, in real time. All of this can be done in the context of a user-supplied background illustration; publication-ready, high-resolution image files can then be exported. Search results can also be saved in a number of abstract file formats for later use elsewhere, including the proteomics standards initiative-molecular interaction format (PSI-MI) molecular interaction standard (9). The protein information pop-up window (Figure 1, bottom) has also been re-designed using the Flash framework and now shows all available 3D structure information for a protein in the context of its domain architecture, which can be browsed interactively along the protein from N- to C-terminus. Apart from PDB entries, the structure information now also includes pre-computed homology models, made available via a collaboration with the SwissModel repository (44).

The current extent of protein–protein association information in STRING is summarized in Figure 2. The majority of associations actually derive from predictions—either from prediction algorithms that are based on analyzing genomic information ('genomic context'-methods) or from transferring associations/interactions between organisms ('interolog'-transfer). Importantly, all associations in STRING are provided with a probabilistic confidence score, which is derived by separately benchmarking groups of associations against the manually curated functional classification scheme of the KEGG database (22). Each score represents a rough estimate of how likely a given association describes a functional linkage between two proteins that is at least as specific as that between an average pair of proteins annotated on the same 'map' or 'pathway'

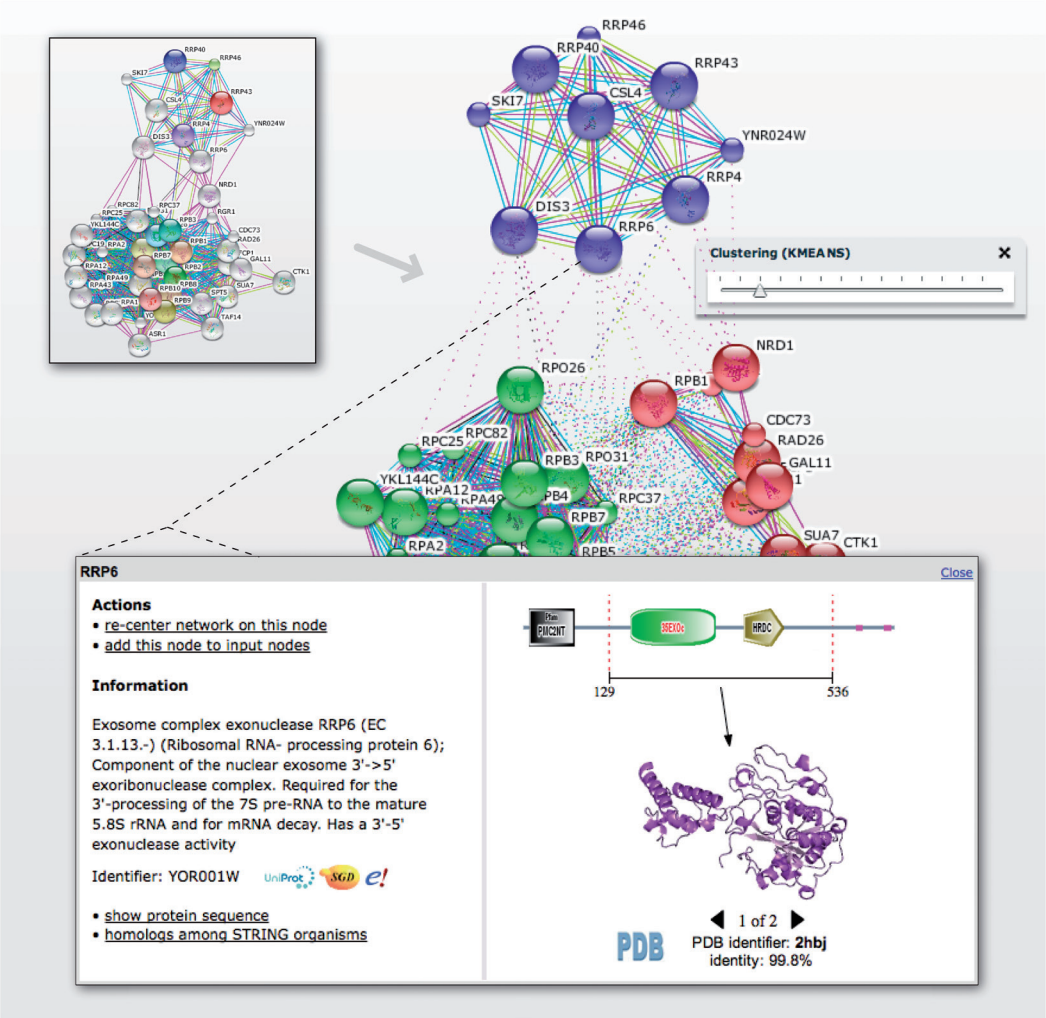


Figure 1. Protein network visualization on the STRING website. The figure shows a composite of two screenshots, illustrating a typical user interaction with STRING (focused on a specific protein network in *Saccharomyces cerevisiae*). Upon querying the database with four yeast proteins, the resource first reports a raw network consisting of the highest scoring interaction partners (upper left corner). This network can then be rearranged and clustered directly in the browser window revealing tightly connected functional modules (arrow). For each interaction (or protein), additional information is accessible via dedicated pop-up windows; the bottom part of the figure shows an exemplary pop-up with the information regarding a specific yeast protein.

in KEGG. The various major sources of interaction/association data in STRING are benchmarked independently; a combined score is computed which indicates higher confidence when more than one type of information supports a given association. All scores and association data in STRING are pre-computed and are also available for wholesale download (free for non-profit institutions). Fully sequenced genomes in STRING are imported from RefSeq (45) and Ensembl (46), as well as from a number of dedicated sites, and are hand-screened for completeness and non-redundancy. For this large space of complete genomes, STRING also stores the results of exhaustive cross-genome homology searches, in order to be able to transfer interactions among organisms. As of version 9.0, this extensive body of protein–protein similarity data is imported from and cross-linked with the Similarity Matrix of Proteins (SIMAP) project (47).

D564 *Nucleic Acids Research, 2011, Vol. 39, Database issue*

	E. coli	*S. cerevisiae*	*H. sapiens*	STRING total
predicted associations				
via gene neighborhood	14,685	0	0	6.9 Mio
via gene fusions	2,930	1,127	4,017	1.3 Mio
via gene co-occurrence	148,420	1,046	7,583	33.0 Mio
via gene co-expression	56,397	58,848	26,278	674,416
known associations				
textmining (co-occurrence)	26,796	70,760	5.5 Mio	7.3 Mio
textmining (natural language processing)	1,000	3,155	226,336	274,214
BIND	368	3,014	2,284	16,741
BioCarta	0	0	12,761	12,761
BioCyc	2,843	809	969	160,462
BioGrid	0	297,276	30,443	365,156
DIP	1,433	12,429	1,088	23,589
Gene Ontology protein complexes	82	6,980	8,052	35,190
HPRD	0	0	18,848	18,848
INTACT	182,33	75,042	28,562	182,524
KEGG	4,965	6,746	38,810	1.9 Mio
MINT	117	66,083	18,839	126,659
NCI Nature / Pathway Interaction DB	0	0	18,909	18,909
PDB	18,586	5,310	2,443	243,434
Reactome	0	83,855	262,482	2.8 Mio
associations transferred from other organisms				
from predicted associations	142,778	39,744	45,890	44.1 Mio
from known associations	21,029	20,152	56,386	12.8 Mio

Figure 2. Association counts and data sources. The table shows the number of pair-wise protein–protein associations processed for STRING (version 8.3), listed separately for three important model organisms as well as for the database as a whole. The associations are counted non-directionally, i.e. protein pairs A–B and B–A are counted only once. Identical associations reported by different sources are counted separately under each source, unless they can be traced to the very same publication record and have been imported from primary interaction databases (in case several such databases agree on an interaction, it is arbitrarily counted for only one of them).

It should be stressed that interactions in STRING are not limited to direct, physical interactions between two proteins. Instead, proteins may also be linked because, for example, they exhibit a genetic interaction or are known to catalyze subsequent steps in a metabolic pathway. Most associations, especially when derived from one of the prediction algorithms, currently can neither be specified with much precision in terms of their mode of interaction, nor in terms of the cellular conditions under which they occur (e.g. development time points, environmental conditions, specific cell types, etc.). Because of this, the fundamental unit stored in STRING is the 'functional association', i.e. the specific and biologically meaningful functional connection between two proteins. Within this definition, STRING aims to uncover the entire space of 'possible' interactions for any fully sequenced organism; it is likely that only a subset of these interactions will be realized in any given cell. The number of interactions stored in STRING has grown considerably over the years and is projected to grow further as more information becomes available. Previous versions of the resource are kept accessible online, such that studies that refer to a given version of STRING can later be reproduced.

Integration with other resources

One central aim of the STRING project is to achieve and maintain cross-connectivity and integration with other public resources in a user-friendly manner. Apart from making the entire SQL database back-end available for download (free for non-profit institutions), this is mainly achieved via the following routes:

First, the database maintains mutual HTML cross-references with a number of widely used websites, including UniProt (48), SMART (49), GeneCards (50) and SwissModelRepository (44). Notably, such cross references do not have to be limited to simple text-based HTML links. Instead, partner websites can embed minimized icon-previews of STRING networks within their own web pages, using the capabilities of STRINGs API interface (as described in the last update) (40). The SMART and SwissModelRepository sites already

use this option, requesting the network preview images—when needed, at run time—based on pre-determined name-space mappings. Such embedded previews do not have to be limited to static images; external sites can also provide pop-up windows for any protein of interest, the content of the pop-up is then provided by STRING [variants of this mechanism are currently used by the resources Reflect (51) and ViralZone (http://expasy.org/viralzone)]. As another new feature of the user interface, permanent URLs can now be retrieved for almost all pages served by STRING—this facilitates cross-linking and archiving and also indexing by search engines and meta-sites.

Second, partner websites can choose to embed the entire STRING website into their own pages (52,53), for example, using HTML inline frames (iframes). A notable example for this is the BioGPS Community Gene Portal System (53); this site provides 'plugins' through which users can connect any number of external websites into freely configurable screen layouts. A STRING plugin has been established at BioGPS; it is currently among the most frequently used plugins there.

Third, users can choose to work with STRING networks from inside the Cytoscape software. Cytoscape is a widely used open-source software framework for network visualization and manipulation (54,55); it can be very flexibly extended, with a rapidly growing number of network-centered manipulation and analysis tools. There are several options for loading STRING data into Cytoscape: users can save a given network from the STRING site to a local file, which can then be opened by Cytoscape (preferably using the PSI-MI format). Users can also query STRING directly from within Cytoscape; this is made possible via a dedicated plugin 'StringWSClient' that exposes much of the STRING query interface, including organism disambiguation. Lastly, the perhaps most important way to query STRING from within Cytoscape is via the 'PSICQUIC' query interface ('PSICQUIC Web Service Universal Client' in Cytoscape). PSICQUIC is a newly developed

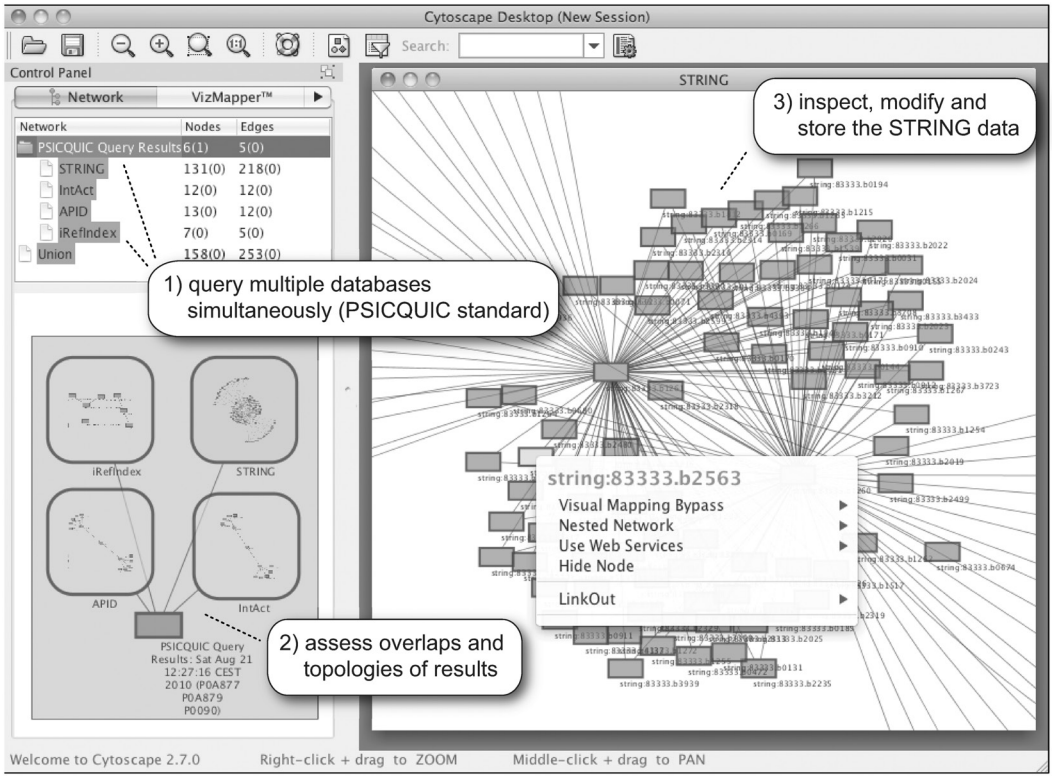


Figure 3. Accessing STRING data from within Cytoscape. Two proteins from *Escherichia coli* were used as queries for the 'PSICQUIC Web Service Universal Client' import-plugin of Cytoscape. Multiple databases have reported hits for these queries (upper left panel); in this case STRING has reported the largest number of hits. The resulting four networks are largely non-overlapping, both in terms of name-spaces as well as in terms of the actual interactors reported. The imported STRING network (right) is shown in detail; it can be used as the basis of further refinement, post-processing and analysis in Cytoscape.

D566 *Nucleic Acids Research, 2011, Vol. 39, Database issue*

standard that allows interaction queries across a growing number of compliant database resources (56); STRING has implemented this standard as of version 8.3 and can thus now be queried directly alongside a number of other resources (Figure 3).

Lastly, a new call-back interface allows STRING to be 'branded' by third-party resources, who may wish to project their own information onto the STRING name space and thereby onto the STRING network data (Figure 4). This allows such resources to take advantage of the extensive user-interface features of STRING, as well as tapping into the existing user base, with very little additional coding effort of their own. This mechanism requires no specific setup on the STRING side—instead, our resource is simply instructed to query the third-party site at runtime, for any additional information that is to be displayed alongside the STRING network. Data updates at the STRING site are usually accommodated automatically, since the name space itself is changed only at the major release updates.

Published use cases

STRING has been used in projects of various scales—both in large, organism-wide studies but also in focused projects that are restricted to a few proteins or to a single pathway only. Studies of the latter type often make use of STRING as a discovery tool, taking advantage of the pre-computed and confidence-scored association predictions that it provides. Examples include the discoveries of a missing enzyme in Bacillothiol biosynthesis in Bacilli (57), of a previously unknown chaperone subunit in Cytochrome C oxidase assembly (58) or of a missing enzyme in uric acid degradation in mammals (59).

Another way to use STRING is to download and extend its relational database schema; this can, for example, be useful for projects dedicated to additional types of information (e.g. small molecule interactors in the case of our partner project STITCH) (60) or for projects wishing to rely on a single source of completely sequenced genomes with associated homology data (e.g. in the case of the gene orthology resource eggNOG) (61). Users not wishing to download and install the entire database schema have the alternative to download compact flat-files; these contain only the actual interaction information or information regarding the interacting proteins themselves (sequences, identifiers, etc.).

A unique strength of STRING lies in its comprehensiveness, albeit at the expense of considerable false-positive rates. Because of this, organism-wide studies represent perhaps the most interesting use cases and they are probably best done when they involve integration of orthogonal data types (since this may allow the noise in both data sets to cancel out). Examples include the filtering and extension of results from large-scale genetic screens (62,63) or the annotation of large groups of proteins having a specific post-translational modification (64). Another intriguing application scenario is to use STRING for search-space reduction in epistasis screens. This is done under the assumption that gene loci showing genetic epistasis should also often show up as functionally linked in STRING. Indeed, this approach has been demonstrated to work on human association mapping data, providing the statistical power to link up loci that show a non-additive effect when mutated together (1,2). Approaches such as this are expected to gain further power, as the information in STRING becomes even more comprehensive and precise in future updates.

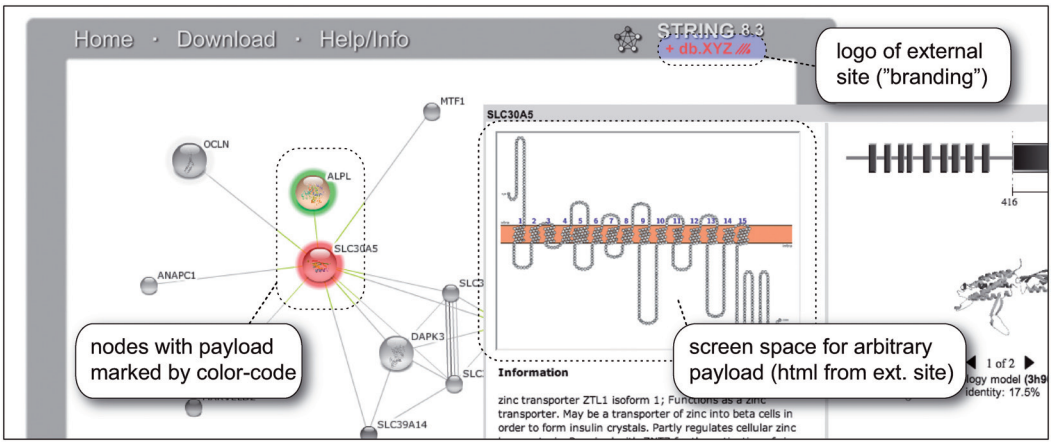


Figure 4. Projecting third-party data onto the STRING web-surface. STRING provides a consistent name space that encompasses genes, genomes, protein and interaction networks, all of which can be easily searched and browsed. These features can now be employed by external web-resources, via a simple call-back mechanism. External resources can provide cross-links to STRING, together with a call-back address capable of serving a simple text-based interface protocol. At run-time, STRING will then automatically call the external site and project arbitrary 'payload' information onto the protein network that is being browsed. The figure shows a fictitious example scenario, served from an in-house test server. As of version 9.0, STRING will also be able to accept protein–protein connections as payload, showing them in a dedicated 'evidence channel' distinct from the seven built-in channels. Implementation details are available in the online documentation.

ACKNOWLEDGEMENTS

The authors wish to thank the PSICQUIC consortium for early access to their standardization effort, and Dr Gary Bader for technical help with the Cytoscape plugin.

FUNDING

STRING is funded by the Swiss Institute of Bioinformatics, by the Novo Nordisk Foundation Center for Protein Research and by the European Molecular Biology Laboratory (EMBL). Funding for open access charges: University of Zurich, through its Research Priority program 'Systems Biology and Functional Genomics'.

Conflict of interest statement. None declared.

REFERENCES

1. Pattin,K.A. and Moore,J.H. (2009) Role for protein-protein interaction databases in human genetics. *Expert Rev. Proteomics*, **6**, 647–659.
2. Emily,M., Mailund,T., Hein,J., Schauser,L. and Schierup,M.H. (2009) Using biological networks to search for interacting loci in genome-wide association studies. *Eur. J. Hum. Genet*., **17**, 1231–1240.
3. Pujol,A., Mosca,R., Farres,J. and Aloy,P. (2010) Unveiling the role of network and systems biology in drug discovery. *Trends Pharmacol. Sci*., **31**, 115–123.
4. Klipp,E., Wade,R.C. and Kummer,U. (2010) Biochemical network-based drug-target prediction. *Curr. Opin. Biotechnol*., **21**, 511–516.
5. Janga,S.C., Diaz-Mejia,J.J. and Moreno-Hagelsieb,G. (2010) Network-based function prediction and interactomics: The case for metabolic enzymes. *Metab. Eng*.
6. Orth,J.D. and Palsson,B.O. (2010) Systematizing the generation of missing metabolic knowledge. *Biotechnol. Bioeng*., **107**, 403–412.
7. Wang,P.I. and Marcotte,E.M. (2010) It's the machine that matters: Predicting gene function and phenotype from protein networks. *J. Proteomics*, **73**, 2277–2289.
8. Lage,K., Karlberg,E.O., Storling,Z.M., Olason,P.I., Pedersen,A.G., Rigina,O., Hinsby,A.M., Tumer,Z., Pociot,F., Tommerup,N. *et al*. (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol*., **25**, 309–316.
9. Hermjakob,H., Montecchi-Palazzi,L., Bader,G., Wojcik,J., Salwinski,L., Ceol,A., Moore,S., Orchard,S., Sarkans,U., von Mering,C. *et al*. (2004) The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data. *Nat. Biotechnol*., **22**, 177–183.
10. Luciano,J.S. (2005) PAX of mind for pathway researchers. *Drug Discov. Today*, **10**, 937–942.
11. Hucka,M., Finney,A., Sauro,H.M., Bolouri,H., Doyle,J.C., Kitano,H., Arkin,A.P., Bornstein,B.J., Bray,D., Cornish-Bowden,A. *et al*. (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **19**, 524–531.
12. Le Novere,N., Hucka,M., Mi,H., Moodie,S., Schreiber,F., Sorokin,A., Demir,E., Wegner,K., Aladjem,M.I., Wimalaratne,S.M. *et al*. (2009) The Systems Biology Graphical Notation. *Nat. Biotechnol*., **27**, 735–741.
13. Lloyd,C.M., Halstead,M.D. and Nielsen,P.F. (2004) CellML: its future, present and past. *Prog. Biophys. Mol. Biol*., **85**, 433–450.
14. Orchard,S., Kerrien,S., Jones,P., Ceol,A., Chatr-Aryamontri,A., Salwinski,L., Nerothin,J. and Hermjakob,H. (2007) Submit your interaction data the IMEx way: a step by step guide to trouble-free deposition. *Proteomics*, **7(Suppl 1)**, 28–34.
15. Salwinski,L., Miller,C.S., Smith,A.J., Pettit,F.K., Bowie,J.U. and Eisenberg,D. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res*., **32**, D449–D451.
16. Aranda,B., Achuthan,P., Alam-Faruque,Y., Armean,I., Bridge,A., Derow,C., Feuermann,M., Ghanbarian,A.T., Kerrien,S., Khadake,J. *et al*. (2010) The IntAct molecular interaction database in 2010. *Nucleic Acids Res*., **38**, D525–D531.
17. Ceol,A., Chatr Aryamontri,A., Licata,L., Peluso,D., Briganti,L., Perfetto,L., Castagnoli,L. and Cesareni,G. (2010) MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res*., **38**, D532–D539.
18. Guldener,U., Munsterkotter,M., Oesterheld,M., Pagel,P., Ruepp,A., Mewes,H.W. and Stumpflen,V. (2006) MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res*., **34**, D436–D441.
19. Chautard,E., Ballut,L., Thierry-Mieg,N. and Ricard-Blum,S. (2009) MatrixDB, a database focused on extracellular protein-protein and protein-carbohydrate interactions. *Bioinformatics*, **25**, 690–691.
20. Goll,J., Rajagopala,S.V., Shiau,S.C., Wu,H., Lamb,B.T. and Uetz,P. (2008) MPIDB: the microbial protein interaction database. *Bioinformatics*, **24**, 1743–1744.
21. Breitkreutz,B.J., Stark,C., Reguly,T., Boucher,L., Breitkreutz,A., Livstone,M., Oughtred,R., Lackner,D.H., Bahler,J., Wood,V. *et al*. (2008) The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res*., **36**, D637–D640.
22. Kanehisa,M., Goto,S., Furumichi,M., Tanabe,M. and Hirakawa,M. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res*., **38**, D355–D360.
23. Matthews,L., Gopinath,G., Gillespie,M., Caudy,M., Croft,D., de Bono,B., Garapati,P., Hemish,J., Hermjakob,H., Jassal,B. *et al*. (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res*., **37**, D619–D622.
24. Karp,P.D., Ouzounis,C.A., Moore-Kochlacs,C., Goldovsky,L., Kaipa,P., Ahren,D., Tsoka,S., Darzentas,N., Kunin,V. and Lopez-Bigas,N. (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res*., **33**, 6083–6089.
25. Keshava Prasad,T.S., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafreen,B., Venugopal,A. *et al*. (2009) Human Protein Reference Database–2009 update. *Nucleic Acids Res*., **37**, D767–D772.
26. Tweedie,S., Ashburner,M., Falls,K., Leyland,P., McQuilton,P., Marygold,S., Millburn,G., Osumi-Sutherland,D., Schroeder,A., Seal,R. *et al*. (2009) FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Res*., **37**, D555–D559.
27. Rodriguez-Esteban,R. (2009) Biomedical text mining and its applications. *PLoS Comput. Biol*., **5**, e1000597.
28. Hoffmann,R. and Valencia,A. (2004) A gene network for navigating the literature. *Nat. Genet*., **36**, 664.
29. Lewis,A.C., Saeed,R. and Deane,C.M. (2010) Predicting protein-protein interactions in the context of protein evolution. *Mol. Biosyst*., **6**, 55–64.
30. Skrabanek,L., Saini,H.K., Bader,G.D. and Enright,A.J. (2008) Computational prediction of protein-protein interactions. *Mol. Biotechnol*., **38**, 1–17.
31. Huynen,M.A., Snel,B., von Mering,C. and Bork,P. (2003) Function prediction and protein networks. *Curr. Opin. Cell Biol*., **15**, 191–198.
32. Valencia,A. and Pazos,F. (2002) Computational methods for the prediction of protein interactions. *Curr. Opin. Struct. Biol*., **12**, 368–373.
33. Hu,Z., Hung,J.H., Wang,Y., Chang,Y.C., Huang,C.L., Huyck,M. and DeLisi,C. (2009) VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology. *Nucleic Acids Res*., **37**, W115–W121.
34. Warde-Farley,D., Donaldson,S.L., Comes,O., Zuberi,K., Badrawi,R., Chao,P., Franz,M., Grouios,C., Kazi,F., Lopes,C.T. *et al*. (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res*., **38(Suppl)**, W214–W220.
35. Kao,H.L. and Gunsalus,K.C. (2008) Browsing multidimensional molecular networks with the generic network browser (N-Browse). *Curr. Protoc. Bioinformatics*, **Chapter 9**, Unit 9 11.

36. Niu,Y., Otasek,D. and Jurisica,I. (2010) Evaluation of linguistic features useful in extraction of interactions from PubMed; application to annotating known, high-throughput and predicted interactions in I2D. *Bioinformatics*, **26**, 111–119.
37. Prieto,C. and De Las Rivas,J. (2006) APID: Agile Protein Interaction DataAnalyzer. *Nucleic Acids Res.*, **34**, W298–W302.
38. Myers,C.L., Chiriac,C. and Troyanskaya,O.G. (2009) Discovering biological networks from diverse functional genomic data. *Methods Mol. Biol.*, **563**, 157–175.
39. Kamburov,A., Wierling,C., Lehrach,H. and Herwig,R. (2009) ConsensusPathDB–a database for integrating human functional interaction networks. *Nucleic Acids Res.*, **37**, D623–D628.
40. Jensen,L.J., Kuhn,M., Stark,M., Chaffron,S., Creevey,C., Muller,J., Doerks,T., Julien,P., Roth,A., Simonovic,M. *et al.* (2009) STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, **37**, D412–D416.
41. Pavlopoulos,G.A., Moschopoulos,C.N., Hooper,S.D., Schneider,R. and Kossida,S. (2009) jClust: a clustering and visualization toolbox. *Bioinformatics*, **25**, 1994–1996.
42. de Hoon,M.J., Imoto,S., Nolan,J. and Miyano,S. (2004) Open source clustering software. *Bioinformatics*, **20**, 1453–1454.
43. Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.
44. Kiefer,F., Arnold,K., Kunzli,M., Bordoli,L. and Schwede,T. (2009) The SWISS-MODEL Repository and associated resources. *Nucleic Acids Res.*, **37**, D387–D392.
45. Pruitt,K.D., Tatusova,T., Klimke,W. and Maglott,D.R. (2009) NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res.*, **37**, D32–D36.
46. Hubbard,T.J., Aken,B.L., Ayling,S., Ballester,B., Beal,K., Bragin,E., Brent,S., Chen,Y., Clapham,P., Clarke,L. *et al.* (2009) Ensembl 2009. *Nucleic Acids Res.*, **37**, D690–D697.
47. Rattei,T., Tischler,P., Gotz,S., Jehl,M.A., Hoser,J., Arnold,R., Conesa,A. and Mewes,H.W. (2010) SIMAP–a comprehensive database of pre-calculated protein sequence similarities, domains, annotations and clusters. *Nucleic Acids Res.*, **38**, D223–D226.
48. Apweiler,R., Martin,M.J., O'Donovan,C., Magrane,M., Alam-Faruque,Y., Antunes,R., Barell,D., Bely,B., Bingley,M., Binns,D. *et al.* (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, **38**, D142–D148.
49. Letunic,I., Doerks,T. and Bork,P. (2009) SMART 6: recent updates and new developments. *Nucleic Acids Res.*, **37**, D229–D232.
50. Safran,M., Dalah,I., Alexander,J., Rosen,N., Iny Stein,T., Shmoish,M., Nativ,N., Bahir,I., Doniger,T., Krug,H. *et al.* (2010) GeneCards Version 3: the human gene integrator. *Database*, **2010**, baq020.
51. Pafilis,E., O'Donoghue,S.I., Jensen,L.J., Horn,H., Kuhn,M., Brown,N.P. and Schneider,R. (2009) Reflect: augmented browsing for the life scientist. *Nat. Biotechnol.*, **27**, 508–510.
52. Liebel,U., Kindler,B. and Pepperkok,R. (2005) Bioinformatic "Harvester": a search engine for genome-wide human, mouse, and rat protein resources. *Methods Enzymol.*, **404**, 19–26.
53. Wu,C., Orozco,C., Boyer,J., Leglise,M., Goodale,J., Batalov,S., Hodge,C.L., Haase,J., Janes,J., Huss,J.W. 3rd *et al.* (2009) BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol.*, **10**, R130.
54. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
55. Cline,M.S., Smoot,M., Cerami,E., Kuchinsky,A., Landys,N., Workman,C., Christmas,R., Avila-Campilo,I., Creech,M., Gross,B. *et al.* (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.*, **2**, 2366–2382.
56. Orchard,S., Albar,J.P., Deutsch,E.W., Eisenacher,M., Binz,P.A. and Hermjakob,H. (2010) implementing data standards: a report on the HUPOPSI workshop September 2009, Toronto, Canada. *Proteomics*, **10**, 1895–1898.
57. Gaballa,A., Newton,G.L., Antelmann,H., Parsonage,D., Upton,H., Rawat,M., Claiborne,A., Fahey,R.C. and Helmann,J.D. (2010) Biosynthesis and functions of bacillithiol, a major low-molecular-weight thiol in Bacilli. *Proc. Natl Acad. Sci. USA*, **107**, 6482–6486.
58. Banci,L., Bertini,I., Ciofi-Baffoni,S., Katsari,E., Katsaros,N., Kubicek,K. and Mangani,S. (2005) A copper(I) protein possibly involved in the assembly of CuA center of bacterial cytochrome c oxidase. *Proc. Natl Acad. Sci. USA*, **102**, 3994–3999.
59. Ramazzina,I., Folli,C., Secchi,A., Berni,R. and Percudani,R. (2006) Completing the uric acid degradation pathway through phylogenetic comparison of whole genomes. *Nat. Chem. Biol.*, **2**, 144–148.
60. Kuhn,M., Szklarczyk,D., Franceschini,A., Campillos,M., von Mering,C., Jensen,L.J., Beyer,A. and Bork,P. (2010) STITCH 2: an interaction network database for small molecules and proteins. *Nucleic Acids Res.*, **38**, D552–D556.
61. Muller,J., Szklarczyk,D., Julien,P., Letunic,I., Roth,A., Kuhn,M., Powell,S., von Mering,C., Doerks,T., Jensen,L.J. *et al.* (2010) eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res.*, **38**, D190–D195.
62. Wang,L., Tu,Z. and Sun,F. (2009) A network-based integrative approach to prioritize reliable hits from multiple genome-wide RNAi screens in Drosophila. *BMC Genomics*, **10**, 220.
63. Mummery-Widmer,J.L., Yamazaki,M., Stoeger,T., Novatchkova,M., Bhalerao,S., Chen,D., Dietzl,G., Dickson,B.J. and Knoblich,J.A. (2009) Genome-wide analysis of Notch signalling in Drosophila by transgenic RNAi. *Nature*, **458**, 987–992.
64. Choudhary,C., Kumar,C., Gnad,F., Nielsen,M.L., Rehman,M., Walther,T.C., Olsen,J.V. and Mann,M. (2009) Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science*, **325**, 834–840.

9.3 RNAi screen of Salmonella invasion shows role of COPI in membrane targeting of cholesterol and Cdc42

9.3.1 Preface

Salmonella Typhimurium is a pathogen and notorious for causing diarrhea. The mechanisms of host cell invasion are not fully understood, but certainly encompass the following steps: *a*) docking of the pathogen at the host membrane; *b*) effector translocation from pathogen to host; *c*) the folding of 'ruffles' by the host, which is induced by the effectors; *d*) host cell invasion by the pathogen; *e*) maturation of the Salmonella-containing vacuole. To gain further insights into these mechanisms the group of Wolf-Dietrich Hardt at ETHZ conducted a siRNA screen comprising 6'978 genes. The knockdown of 298 of these genes resulted in either an enhancement or a reduction of invasion efficiency by a factor of more than 1.5. We were approached to perform a cluster analysis of these candidate hits, as this might reveal functional modules, which take part in the invasion process. Using hierarchical clustering, we detected two clusters that showed a significant enrichment of protein-protein interactions as annotated in the STRING database. The first cluster contained a network around known regulators of membrane ruffling (Cdc42, NckAP1) and the second contained genes belonging to the coatomer complex, which is known to play an important role in membrane translocation. Together with their functionally already described members, the genes contained within these clusters provide a suitable basis for future research, which will further elucidate the process of bacterial host cell invasion.

9.3.2 Molecular Systems Biology, 2011

Molecular Systems Biology 7; Article number 474; doi:10.1038/msb.2011.7
Citation: *Molecular Systems Biology* 7:474

www.molecularsystemsbiology.com

molecular systems biology

RNAi screen of *Salmonella* invasion shows role of COPI in membrane targeting of cholesterol and Cdc42

Benjamin Misselwitz[1,4], Sabrina Dilling[1,4], Pascale Vonaesch[1,4], Raphael Sacher[2], Berend Snijder[2], Markus Schlumberger[1], Samuel Rout[1], Manuel Stark[3], Christian von Mering[3], Lucas Pelkmans[2,3] and Wolf-Dietrich Hardt[1,*]

[1] Institute of Microbiology, D-BIOL, ETH Zürich, Zürich, Switzerland, [2] Institute of Molecular Systems Biology, D-BIOL, ETH Zürich, Zürich, Switzerland and [3] Institute of Molecular Life Sciences, University of Zürich, Zürich, Switzerland
[4] These authors contributed equally to this work
* Corresponding author. Institute of Microbiology, D-BIOL, ETH Zürich, Wolfgang-Pauli-Str. 10, Zürich 8093, Switzerland.
Tel.: +41 44 632 5143; Fax: +41 44 632 1129; E-mail: hardt@micro.biol.ethz.ch

Received 8.10.10; accepted 3.2.11

The pathogen *Salmonella* Typhimurium is a common cause of diarrhea and invades the gut tissue by injecting a cocktail of virulence factors into epithelial cells, triggering actin rearrangements, membrane ruffling and pathogen entry. One of these factors is SopE, a G-nucleotide exchange factor for the host cellular Rho GTPases Rac1 and Cdc42. How SopE mediates cellular invasion is incompletely understood. Using genome-scale RNAi screening we identified 72 known and novel host cell proteins affecting SopE-mediated entry. Follow-up assays assigned these 'hits' to particular steps of the invasion process; i.e., binding, effector injection, membrane ruffling, membrane closure and maturation of the *Salmonella*-containing vacuole. Depletion of the COPI complex revealed a unique effect on virulence factor injection and membrane ruffling. Both effects are attributable to mislocalization of cholesterol, sphingolipids, Rac1 and Cdc42 away from the plasma membrane into a large intracellular compartment. Equivalent results were obtained with the vesicular stomatitis virus. Therefore, COPI-facilitated maintenance of lipids may represent a novel, unifying mechanism essential for a wide range of pathogens, offering opportunities for designing new drugs.
Molecular Systems Biology **7:** 474; published online 15 March 2011; doi:10.1038/msb.2011.7
Subject Categories: membranes & transport; microbiology & pathogens
Keywords: coatomer; HeLa; *Salmonella*; siRNA; systems biology

Introduction

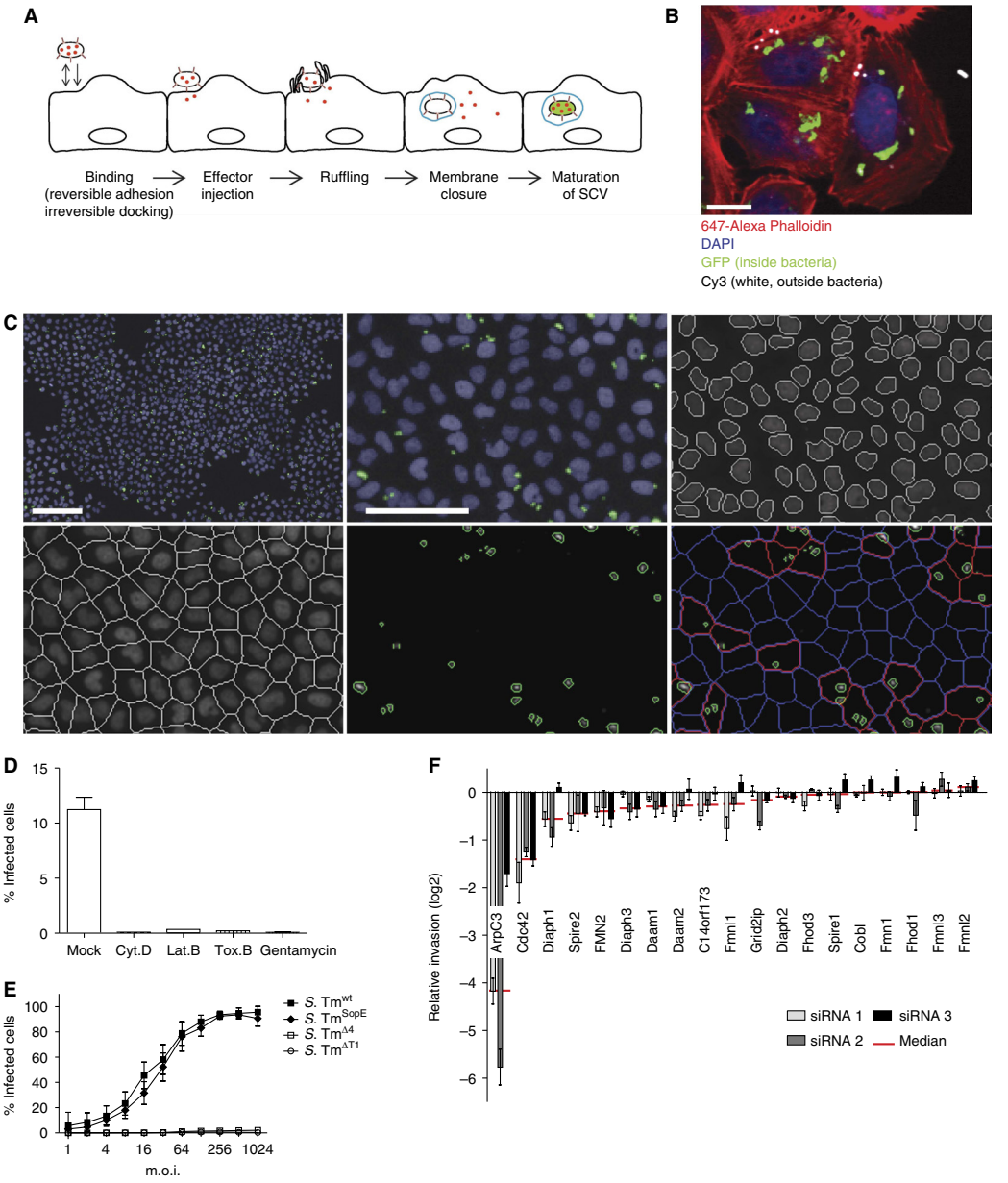
Salmonella enterica subspecies 1 serovar Typhimurium (*S*. Typhimurium or *S*. Tm in this paper) is a common cause of diarrhea in humans. Central to its pathogenicity is the ability to invade gut epithelial cells (Patel and Galan, 2005; Schlumberger and Hardt, 2006; McGhie *et al*, 2009). This process is incompletely understood and requires an intricate interplay of *S*. Typhimurium virulence factors and numerous host cell factors. The whole range of host cell factors involved in the invasion process has not been elucidated.

In order to invade epithelial cells, *S*. Typhimurium binds to the host cell surface. This process can involve reversible adhesion (e.g., via fimbriae) and irreversible docking via the Type III secretion system 1 (T1; Misselwitz *et al*, 2011). Then, T1 acts as a molecular syringe to inject virulence factors, so called effectors (Figure 1A). Four key effectors, SopE, SopE2, SopB and SipA, can trigger actin polymerization and mediate epithelial cell invasion in a functionally overlapping manner (Norris *et al*, 1998; Zhou *et al*, 1999; Schlumberger and Hardt, 2006). Among the key effectors, SopE, a G-nucleotide exchange factor for the Rho GTPases Rac1 and Cdc42 (Hardt *et al*, 1998; Rudolph *et al*, 1999; Friebel *et al*, 2001), is the most potent trigger of invasion. Both Rho GTPases signal to the Arp2/3 complex, a powerful activator for actin polymerization (Goley and Welch, 2006). Actin polymerization leads to the formation of pronounced and characteristic ruffles on the cellular surface (Finlay *et al*, 1991) facilitating invasion. Once inside the host cell, *S*. Typhimurium is enclosed in a vacuole which matures, acquires late endosome markers and positions itself close to the nucleus (Guignot *et al*, 2004; Marsman *et al*, 2004). Inside this '*Salmonella*-containing vacuole' (SCV), *S*. Tm expresses a second set of virulence factors encoded on the '*Salmonella* pathogenicity island 2' (SPI-2) (Schlumberger and Hardt, 2006).

While tremendous progress has been achieved toward understanding the invasion of *S*. Typhimurium into host cells, the picture is far from complete. Importantly, host cell factors

required for *S.* Typhimurium invasion have never been analyzed in a comprehensive and unbiased manner. RNAi screening has recently been introduced to study host–pathogen interactions (Ramet *et al*, 2002; Agaisse *et al*, 2005; Cheng *et al*, 2005; Pelkmans *et al*, 2005; Philips *et al*, 2005; Derre *et al*, 2007; Kuijl *et al*, 2007; Cherry, 2008; Elwell *et al*, 2008; Chong *et al*, 2009; Prudencio and Lehmann, 2009; Hirsch, 2010). This systematic approach has significantly advanced our understanding of the respective molecular infection processes.

To better understand the mechanism of *S.* Typhimurium entry, we performed a genome-scale RNAi screen for host



proteins affecting SopE-mediated invasion into HeLa cells. Our experiments identified actin cytoskeletal regulators and proteins affecting host cell invasion, which were not previously implicated in this process. Systematic follow-up assays revealed a novel functional link between the coatomer I (COPI) complex, the maintenance of the membrane lipid composition and the capacity of *S.* Typhimurium to manipulate its host cell. This novel mechanism might be of general importance for numerous bacterial and viral pathogens.

Results

An automated assay measuring *S.* Typhimurium invasion into host cells

For large-scale screening of *S.* Typhimurium invasion, we used a modified gentamycin protection assay. In our assay, HeLa cells were infected for 20 min with wild-type *S.* Typhimurium (*S.* Tm^{wt} (pM975); Supplementary Table SI) expressing GFP under a SPI-2 promotor (*pssaG*). This GFP reporter is not expressed by extracellular bacteria, but strongly induced when the pathogen resides within the SCV. After infection, medium containing gentamycin (kills all extracellular bacteria) was added, followed by a 4 h incubation step allowing for GFP expression and maturation (Schlumberger *et al*, 2007). Figure 1B shows specific GFP expression by intracellular *S.* Tm^{wt} (pM975; 'green') but not by bacteria remaining extracellular ('white'). The combination of gentamycin protection and SPI-2-driven GFP expression resulted in a specific and bright fluorescence signal of the intracellular bacteria.

To establish a genome-wide screening system, we used automated microscopy and developed an automated algorithm to identify and enumerate cells successfully invaded by *S.* Typhimurium. The assay determines the fraction of infected cells, as defined by the presence of at least one GFP-expressing intracellular bacterium (Figure 1C, bottom right panel, red). Preincubation of cells before the infection with medium containing gentamycin reduced the fraction of infected cells 250- to 1000-fold (Figure 1D), confirming the high specificity of the assay. In addition, toxins known to inhibit *S.* Typhimurium invasion showed similar effects: Cytochalasin D and Latrunculin B, which disrupt the actin cytoskeleton, and Toxin B from *Clostridium difficile*, which inactivates Rho GTPases, reduced the fraction of infected cells by at least 90%. These observations demonstrate that disruption of important host cell signaling pathways yields pronounced signals in this assay, validating the approach.

It should be noted that SPI-2 expression does not occur immediately after *Salmonella* entry but requires a maturation step of the SCV. Therefore, the modified gentamycin protection assay would be sensitive to perturbations of *S.* Typhimurium entry, the first steps of its intracellular life cycle, as well as SCV fusion with the lysosome and can be considered as a global assay probing *Salmonella* entry and SCV maturation.

Testing host cell genes for effects on *S.* Typhimurium invasion

While a strain lacking the four key effectors SipA, SopE, SopE2 and SopB (*S.* $Tm^{\Delta 4}$ (pM975)) showed only marginal invasion, an isogenic mutant lacking three key effectors with only SopE remaining (*S.* Tm^{SopE} (pM975)) invaded almost as efficiently as *S.* Tm^{wt} (pM975) (Figure 1E). As expected, invasion of the mutant without the Type III secretion system 1 (*S.* $Tm^{\Delta T1}$ (pM975)) was negligible. On the basis of these observations, we decided to use *S.* Tm^{SopE} (pM975) for our screen. This circumvented any issues arising from the presence of SopE2, SipA or SopB, as these functionally overlapping effectors might mask phenotypes specific for SopE-induced host cell invasion.

To validate the assay, we screened a small targeted siRNA library systematically depleting known actin nucleators, including formins, the p21 subunit of the Arp2/3 complex (ArpC3) as well as the Rho GTPase Cdc42 (Cdc42). Cells were preincubated with siRNAs for 3 days and *S.* Tm^{SopE} (pM975) invasion was analyzed by automated microscopy. Depletion of ArpC3 reduced invasion by ~95% (log2 median=−4.2; Figure 1F). A similar effect was observed after depletion of Cdc42 (70%, log2 median=−1.4). Depletion of Diaph1, Spire2 and Fmn2 resulted in only 32%, 26% and 25% reduced invasion, respectively, and all other formins yielded even weaker or no detectable effects on invasion at all. These data are in line with the current model of *S.* Typhimurium host cell invasion, which implicates that the activation of Rho GTPases by SopE and subsequent activation of the Arp2/3 complex are essential for SopE-mediated invasion (Schlumberger and Hardt, 2006). In contrast, individual formins contribute much less to *S.* Typhimurium entry. Overall, this experiment confirmed that the image-based invasion assay is well suited and sufficiently robust for performing a genome-scale siRNA screen.

Figure 1 Establishment of an automated assay to analyze *S.* Typhimurium invasion. (**A**) Overview showing the invasion process of *S.* Typhimurium divided into five major steps: (i) during the binding step, the bacteria attach to the cellular surface by reversible adhesion or irreversible docking; (ii) T1 is used as a molecular syringe to inject effectors (shown in red) into the eukaryotic cell; (iii) these effectors in turn induce membrane ruffling; (iv) subsequently the cellular membrane encloses a bacterium (membrane closure), thereby producing a *Salmonella*-containing vacuole (SCV, shown in blue); (v) after a maturation step, *S.* Tm genes important for intracellular survival are induced (green). (**B**) Fluorescence image showing GFP expression of *S.* Tm^{wt} (pM975) only after invasion into HeLa cells (green=inside bacteria, red=actin, blue=DAPI, white=outside bacteria; scale bar=20 μm). (**C**) Automated image analysis strategy: *S.* Tm^{wt} (pM975) infection of HeLa cells followed by the acquisition of nuclei (blue) and bacterial spots (green) using an automated microscope with a ×10 objective. Images were analyzed using CellProfiler as follows: recognition of nuclei, definition of cells, identification of bacterial spots and the allocation of these spots to cells (red outline=infected cell, blue outline=non-infected cell; scale bar whole image=100 μm, detailed image=50 μm). (**D**) Verification of the automated assay testing inhibitors of *Salmonella* invasion. HeLa cells were infected with *S.* Tm^{wt} (pM975) and analyzed as described in (C). Pretreatment of HeLa cells with the inhibitors Cytochalasin D (Cyt. D), Latrunculin B (Lat. B), Toxin B (Tox. B) or the antibiotic gentamycin prevents invasion. (**E**) Invasion efficiencies of various *Salmonella* strains into HeLa cells analyzed by the automated assay showing *S.* Tm^{SopE} (pM975) invasion being as efficient as *S.* Tm^{wt} (pM975). (**F**) Verification of the automated assay using siRNAs directed against different actin polymerization regulators. Depletion of ArpC3 and Cdc42 reduces *S.* Tm^{SopE} (pM975) invasion (red line=median of three siRNAs tested for each gene; log2 relative invasion=% infected cells with siRNA treatment divided by the median of % infected cells treated with control siRNA).

A genome-scale screen for host cell proteins affecting SopE-mediated invasion

To identify host cell genes affecting SopE-mediated invasion on a genome scale, we used the 'druggable genome' siRNA library (Version 2.0, Qiagen; Supplementary Table SII) covering 6978 human genes. The siRNAs were transferred into 384-well dishes. Each dish included identical sets of controls for siRNA transfection (Eg5 and Plk1 reduce cell number significantly) and positive controls, including ArpC3 (known negative effect on invasion) or gentamycin (kills bacteria before invasion). These controls allowed direct comparison of the data between the different plates and provided essential quality controls.

SiRNA-transfected HeLa cells were infected with *S.* Tm^{SopE} (pM975) at an m.o.i. of 64. In this screen, each gene was tested with three independent siRNAs and each test was performed in triplicates. The invasion efficiency was analyzed by the modified gentamycin protection assay. For each plate, the values were normalized to the median of all siRNAs on this plate. In addition, *z*-score correction was performed, yielding 361 candidate hits from the uncorrected and 190 candidate hits from the *z*-score corrected data (for details see Materials and methods). 'Hit' genes encoding central subunits of the Arp2/3 complex (Actr2, Actr3) and Cdc42 showed strong inhibitory effects, thus validating our approach (Figure 2A and B). On the basis of the lists of candidate hits and possible interaction partners not present in the original library, we assembled a new library targeting 298 genes with four siRNAs per gene (Supplementary Table SIII). This library also included siRNAs to control for transfection efficiency (Eg5, Plk1 killing cells) and effects on *S.* Tm invasion (ArpC3, Cdc42, Cfl1).

The library was tested in a confirmatory screen for SopE-mediated invasion. *S.* Tm invasion was normalized using the median of 12 control siRNAs targeting genes without effects on *S.* Tm invasion in the druggable genome screen (Materials and methods). This allowed normalization in spite of the highly biased nature of the confirmatory library (i.e., containing hits of the primary screen), which prohibited reliable threshold calculation based on the screened library, itself. Therefore, a hit was arbitrarily defined as a gene displaying a log2 of the median of the 4 siRNAs $\leqslant -0.5$ or $\geqslant 0.3$. The confirmatory screen thus validated 72 hits (Figure 2A, insert; Supplementary Table SIII). A subset of the results is represented in Figure 2C and depletion efficiencies of these hits were verified (Supplementary Figure 1). Some hits of the primary screen could not be confirmed. This could be explained by off-target effects of the oligos used in the initial screen. Additionally, in the rescreen four instead of three oligos per gene were tested and some of the oligos were newly designed by the company. These slightly changed conditions could explain the different results obtained in the initial and the confirmatory screen.

The list of confirmed hits included important regulators of the actin cytoskeleton previously implicated in *S.* Typhimurium host cell invasion (Figure 2B and C), as the heptameric Arp2/3 complex of which both subunits tested in the druggable genome screen showed strong inhibitory effects. Other examples include Cdc42 (Chen *et al*, 1996) and the Nck-associated protein 1 (Nap1, *nckp1*; Shi *et al*, 2005; Hanisch *et al*, 2010), a component of the Wave complex linking Rho GTPase activation to the

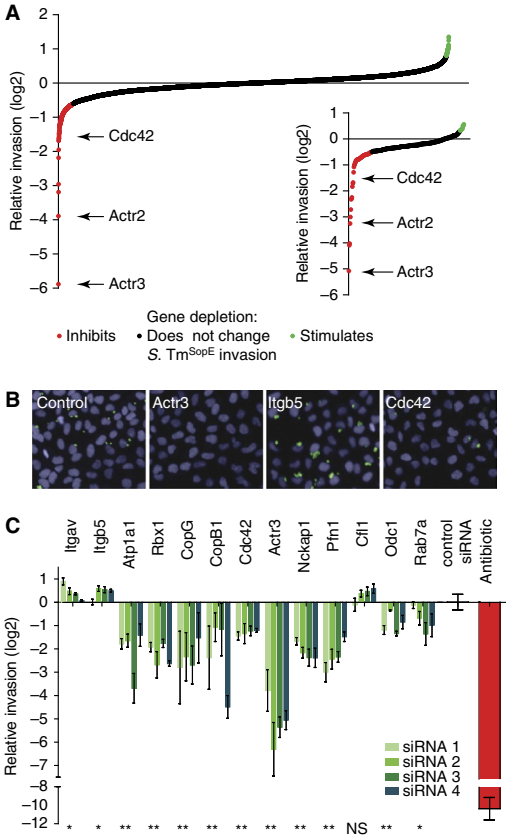


Figure 2 Genome-scale siRNA screen reveals host cell factors required for *Salmonella* invasion. (**A**) Overview of relative invasion of *S.* Tm^{SopE} (pM975) into HeLa cells transfected with siRNA of the druggable genome library. Values represent the median of three siRNAs per gene. Median values above and below 1.5 times of the interquartile range were defined as positive (green) and negative (red) hits, respectively. A confirmatory screen with 298 selected genes approved 72 hits from the genome-scale screen (inserted small graph). Positive (green) and negative (red) hits were determined with median values above 0.3 and below −0.5, respectively. (**B**) Example images of indicated hits from the screen demonstrating reduced invasion of *S.* Tm^{SopE} (pM975) for Actr3- and Cdc42-depleted cells and increased invasion in the absence of Itgb5 (blue=nuclei, green=invaded bacteria; scale bar=50 μm). (**C**) Relative invasion of *S.* Tm^{SopE} (pM975) for selected hits from both screens (**P*-value <0.1, ***P*-value <0.05, NS=not significant, Mann–Whitney *U*-test).

activation of the Arp2/3 complex (Goley and Welch, 2006). In addition, we identified Profilin 1 (Pfn1), a well-characterized actin binding protein which delivers actin monomers to sites of actin polymerization (Pollard and Cooper, 2009), that has not been studied before in the context of *S.* Typhimurium invasion. Hits stimulating *S.* Tm^{SopE} (pM975) invasion efficiency included adenylyl cyclase-associated protein 1 (Cap1; Balcer *et al*, 2003), which mediates the breakdown of actin fibers and can affect *S.* Typhimurium host cell invasion (Maciver and Hussey, 2002; McGhie *et al*, 2004; Paavilainen *et al*, 2004). Therefore, these

hits confirm and extend the current model of SopE-mediated host cell invasion. The screen also identified numerous novel genes not previously linked to *S*. Typhimurium host cell invasion. These hits include the sodium potassium ATPase 1 (Atp1a1; Kaplan, 2002), the ring box protein 1 (Rbx1), an ubiquitin E3 ligase and essential partner for most of the proteins of the Cullin family (Petroski and Deshaies, 2005) and subunits of the heptameric COPI complex. Two subunits of the COPI complex were present in the druggable genome library of which COPB1 was missed. However, the depletion of all five subunits in the confirmatory screen reduced *Salmonella* invasion. Coatomer I is implicated in retrograde transport of vesicles cycling between the Golgi apparatus and the endoplasmic reticulum (Lee *et al*, 2002; Beck *et al*, 2009) and in anterograde transport of some proteins (Pepperkok *et al*, 1993; Orci *et al*, 1997). Two integrins, Itgβ5 (Itgb5) and Itgαv (Itgav; Shimaoka and Springer, 2003), were identified as strong invasion-stimulating hits (Figure 2C). To the best of our knowledge, the proteasome complex has not been implicated in *Salmonella* invasion before. All seven α-subunits and five of seven β-subunits were present in the genome-scale library, of which two α- and two β-subunits were identified as hits, all of which could be confirmed. The detailed composition of the mediator complex implicated in protein splicing in various cell types is still controversial (Conaway *et al*, 2005) but seems to include more than 20 proteins, of which seven were present in the initial library. Only Med4 was identified as a hit. It remains unclear whether the remaining subunits were missed due to experimental noise or whether an effect (if any) on *Salmonella* invasion is restricted to Med4. Taken together, host factors important for SopE-mediated *S*. Tm invasion comprise a surprising variety of cellular components and are not limited to well-established actin-regulating proteins.

Follow-up screen of candidate hits affecting host cell binding

The modified gentamycin protection assay used in the screen measures the presence of *S*. Tm in a mature SCV and could thus identify genes affecting any step of the invasion process. In order to assign the hits to particular steps and enable identification of functional links between the novel hits, we developed step-specific secondary assays addressing *S*. Tm binding, effector injection, cellular ruffling and membrane closure.

Binding to the host cell is the first step of *S*. Typhimurium invasion. Under the conditions used in the screen, binding is mediated mainly by the T1 system itself (Lara-Tejero and Galan, 2009; Misselwitz *et al*, 2011). In order to uncouple binding from the subsequent steps of the invasion process, we utilized the isogenic non-invasive strain *S*. $Tm^{\Delta 4}$ (Supplementary Table SI). This strain encodes a fully functional T1 system, attaches via T1 to host cells and can efficiently insert the T1 translocon conduit into the host cell membrane. However, it lacks the four key effector proteins SipA, SopE, SopE2 and SopB and is thus incapable of triggering the subsequent steps of the invasion process, i.e., membrane ruffling and invasion. To synchronize the binding process for all cells of a well, we chose a short incubation time (6 min) and an m.o.i. of 82. The unbound bacteria were washed off, the cells were fixed, and we stained the nuclei with DAPI and the bound bacteria with an anti-LPS antibody (Materials and methods). Automated microscopy and automated image analysis were adapted to enumerate cells carrying bound bacteria on the surface. Strikingly, mitotic cells always carried much higher numbers of bound bacteria than non-mitotic cells (Figure 3A and B, top panel). Cells neighboring mitotic cells often displayed this phenotype as well. The reason for this increased binding phenotype is not understood. Nevertheless, to increase the sensitivity of our assay, mitotic cells and their direct neighbors were excluded from the analysis (Misselwitz *et al*, 2010).

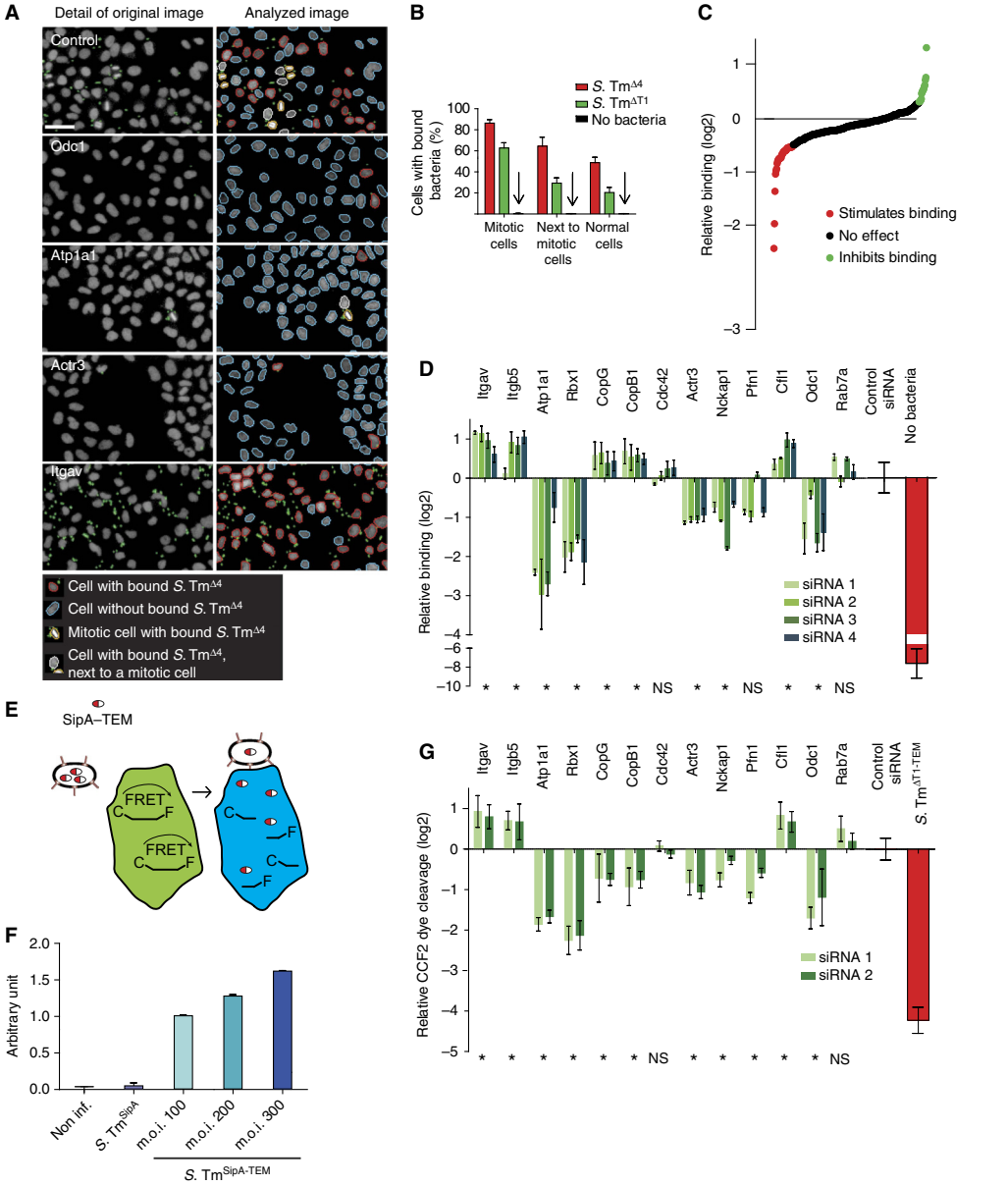
HeLa cells were seeded in 96-well dishes and transfected with the siRNA library for the 298 candidate hits (four siRNAs per gene), including siRNAs for quality control, i.e., Eg5, Plk1 (transfection controls), ArpC3, Cdc42, Cfl1 (strong effects on infection) and the 12 siRNAs without detectable effect for normalization. Binding was normalized to the control siRNAs and a binding hit was defined as a gene displaying a log2 of the median of four siRNAs $\leqslant -0.5$ or $\geqslant 0.3$. By these criteria, 15 hits displayed increased binding and 38 hits displayed decreased binding (Figure 3C; Supplementary Table SIII). Several examples are shown in Figure 3D. By comparing the binding phenotypes with the invasion hits, we made two general observations:

(i) Numerous invasion hits were also identified as binding hits (compare Figures 2C and 3D; Supplementary Table SII). These included Atp1a1 and Rbx1, as well as numerous actin regulators (Actr3, Pfn1, Nckap1) and several integrins (Itgb5, Itgav). In these cases, the invasion phenotype could for some genes partially, for others completely be assigned to the binding step. These binding hits might affect host cellular membrane stiffness, surface charges/hydrophobicity or the binding site/receptor availability on the host cell surface.
(ii) A significant number of invasion hits including Cdc42 and Rab7a, a small GTPase involved in vesicular trafficking did not show reduced binding; for components of the COPI complex (CopB1, CopG), binding was even increased. These genes must affect later steps of the invasion process.

The effector injection assay reveals a specific role of the COPI complex

Upon binding, *S*. Typhimurium must inject T1 effector proteins into the host cell (Figure 1A). Before host cell contact, the T1 system is preassembled, but inactive. Upon binding, a special conduit termed 'translocon' is inserted into the host cell membrane, thus activating the T1 system and initiating effector protein injection into the host cell. Host signals activating the T1 system are not well understood. Some hits identified in the invasion screen might affect this step of the infection process. To measure effector injection, we used an assay based on a fusion protein between β-lactamase and the T1 effector SipA (Figure 3E) (Charpentier and Oswald, 2004; Schlumberger *et al*, 2007). After injection into cells, the β-lactamase part of the fusion protein is able to cleave the fluorescent dye CCF2 present within the cell, thereby changing

its fluorescence properties. As cellular ruffles and invasion might change the efficiency of effector translocation, we chose the strain *S.* Tm^{SipA} for this assay (Supplementary Table SI). This strain carried a fusion protein of SipA and β-lactamase (*S.* $Tm^{SipA–TEM}$), but lacks SopE, SopE2 and SopB, does not trigger ruffling and invades only very slowly and less efficiently into HeLa cells (Schlumberger *et al*, 2005 and data not shown). Infection of CCF2-loaded HeLa cells with *S.* $Tm^{SipA–TEM}$ resulted in a pronounced change in fluorescence (Figure 3F). In contrast, *S.* Tm^{SipA} lacking β-lactamase did not induce any change in fluorescence, confirming the specificity of this assay.



Using this effector injection assay, we focused on a 90-gene subset of the 298 candidate hits and analyzed only the two siRNAs per gene, which had yielded the strongest signals in the invasion assay (Figure 2C). This subset of genes included 72 confirmed hits of the invasion assay as well as genes with effects close to our arbitrary threshold, which together could be conveniently tested on three 96-well plates. HeLa cells were seeded and transfected with siRNA as described for the confirmatory screen (compare Figure 2). The fluorescence signal of the injection assay was normalized to the 12 control siRNAs. Several examples of the injection screen are shown in Figure 3G (compare Supplementary Table SIII). A first comparison with the data of the binding screen (Figure 3A–D) allowed three global conclusions:

(i) Depletion of subunits of the COPI complex strongly impaired effector translocation. In contrast, the depleted cells showed increased binding (Figure 3D). Thus, lack of the COPI complex seems to specifically inhibit effector translocation (Figure 3G).
(ii) All remaining binding hits also displayed injection phenotypes. Depletion of some proteins (e.g., Itgav, Itgb5, Cfl1) enhanced binding and injection, whereas depletion of others (e.g., Atp1a1, Rbx1, Actr3, Nckap1, Odc1) reduced both phenotypes. In these cases, the altered binding efficiency was sufficient to explain the altered injection efficiency.
(iii) A number of invasion hits showed neither reduced injection nor binding. These included the Rho GTPase Cdc42 and Rab7a. These genes must affect later steps of the invasion process.

The ruffling assay yielded a specific phenotype for Cdc42

Upon binding and effector protein injection, HeLa cells respond within minutes by pronounced actin rearrangements and by forming membrane ruffles. SopE is a key effector protein triggering actin rearrangements and membrane ruffles. To screen for ruffling phenotypes, HeLa cells were infected for 6 min at an m.o.i. of 250, fixed and nuclear DNA and the actin cytoskeleton were stained. By fluorescence microscopy, we observed that *S.* Tm^{SopE} triggered pronounced ruffles, whereas the isogenic control strain *S.* $Tm^{\Delta 4}$ lacking SopE, SopE2, SopB and SipA did not change the appearance of the cells (Figure 4A, top left panels).

To analyze the ruffling efficiency in a quantitative manner, an automated microscopy-based assay for the identification of ruffling cells was developed. Starting from images acquired by automated microscopy, nuclei and cells were identified using the image analysis software CellProfiler (Carpenter *et al*, 2006). With the recently released machine learning tool, Enhanced CellClassifier (Misselwitz *et al*, 2010), sets of ruffling and non-ruffling cells were generated to train a support vector machine (SVM) model algorithm (Cortes and Vapnik, 1995) for automatic recognition of ruffling and non-ruffling cells. Representative results of the automatic classification are shown in the lower panels of Figure 4A.

To screen for genes affecting the ruffling step, HeLa cells were seeded in 96-well plates and transfected with the siRNA library for the 298 candidate hits, or control siRNAs as described above and were infected with *S.* Tm^{SopE}. Ruffling was normalized using the control siRNAs and a ruffling hit was defined as a gene displaying a log2 of relative ruffling $\leqslant -0.5$ or $\geqslant 0.3$. Using this cutoff, 29 genes displayed increased ruffling and 52 genes displayed decreased ruffling (Supplementary Table SIII). Several examples are shown in Figure 4B. We made the following observations:

(i) Some injection hits showed equivalent phenotypes in the ruffling assay, e.g., reduced (e.g., Odc1) or enhanced ruffling (e.g., Itgb5, Itgav, Cfl1; compare Figures 3G and 4B).
(ii) For other hits (including CopB, CopG, Atp1a1, Actr3, Nckap1, Pfn1), the phenotype was even stronger in the ruffling assay than in the injection assay, suggesting an additional role for these genes at this step.
(iii) A few ruffling hits (including Cdc42) had not shown a significant effect in binding or injection assays (compare Figures 3D, G and 4B). Therefore, these hits specifically affected the ruffling step.
(iv) Other invasion hits such as Rab7a did not show a significant phenotype in the ruffling assay, suggesting a role in later steps of the invasion process.

An assay for hits affecting membrane closure

Finally, the actual entry step of *S.* Typhimurium into the host cell was investigated. As the modified gentamycin protection assay (Figures 1 and 2) measures *S.* Tm entry only after SCV maturation, an assay was developed which measures the membrane fusion event completing pathogen entry into the

Figure 3 Host cell factors affect *Salmonella* binding and effector injection. (**A**) Establishment of the binding assay for indicated siRNA-transfected HeLa cells (left) and corresponding image analysis (right). To analyze binding independent of cell cycle state, mitotic cells and their neighbors (showing an increased binding phenotype; indicated by yellow/white border) were excluded from the analysis using Enhanced CellClassifier. The remaining nuclei either had bound bacteria or not (nuclei with red/blue border; gray=nuclei, green=*S.* $Tm^{\Delta 4}$, scale bar=50 μm). (**B**) Differential quantification of *S.* $Tm^{\Delta 4}$ and *S.* $Tm^{\Delta T1}$ binding for mitotic cells, neighbors of mitotic cells and other cells. Each bar shows the median and standard deviation from 72 wells of two independent experiments. (**C**) Binding efficiency of *S.* $Tm^{\Delta 4}$ onto HeLa cells transfected with the confirmatory siRNA library. Data are displayed as log2 relative binding corresponding to the percentage of cells with bound bacteria of siRNA-treated cells and control siRNA-treated cells. (**D**) Binding efficiency of *S.* $Tm^{\Delta 4}$ for selected genes. The depletion of Atp1A1 and Rbx1 strongly inhibits binding of *S.* $Tm^{\Delta 4}$, whereas the depletion of Itgav and Itgb5 stimulates adherence to HeLa cells. (**E**) Scheme of the effector injection assay. HeLa cells were loaded with CCF2-AM (green) and infected with *Salmonella* carrying a fusion protein of SipA with β-lactamase (SipA–TEM). Translocated β-lactamase mediates cleavage of CCF2-AM inducing a shift from green to blue fluorescence. Effector translocation is defined as the ratio between blue (460 nm) and green (435 nm) fluorescence. (**F**) Validation of the effector injection assay: HeLa cells were infected with *S.* Tm^{SipA} or *S.* $Tm^{SipA\text{-}TEM}$, an increase of effector injection signal (arbitrary units) was observed with increasing m.o.i. of the bacteria. (**G**) Effector injection analysis of selected genes. HeLa cells transfected with siRNAs directed against *copG* and *copB1* and infected with *S.* $Tm^{SipA\text{-}TEM}$ showed a twofold reduction in effector injection efficiency. The ratio between blue and green fluorescence was normalized to control siRNAs and is displayed as log2 of the relative translocation. Two siRNAs per gene have been analyzed (*$P<0.1$, NS=not significant).

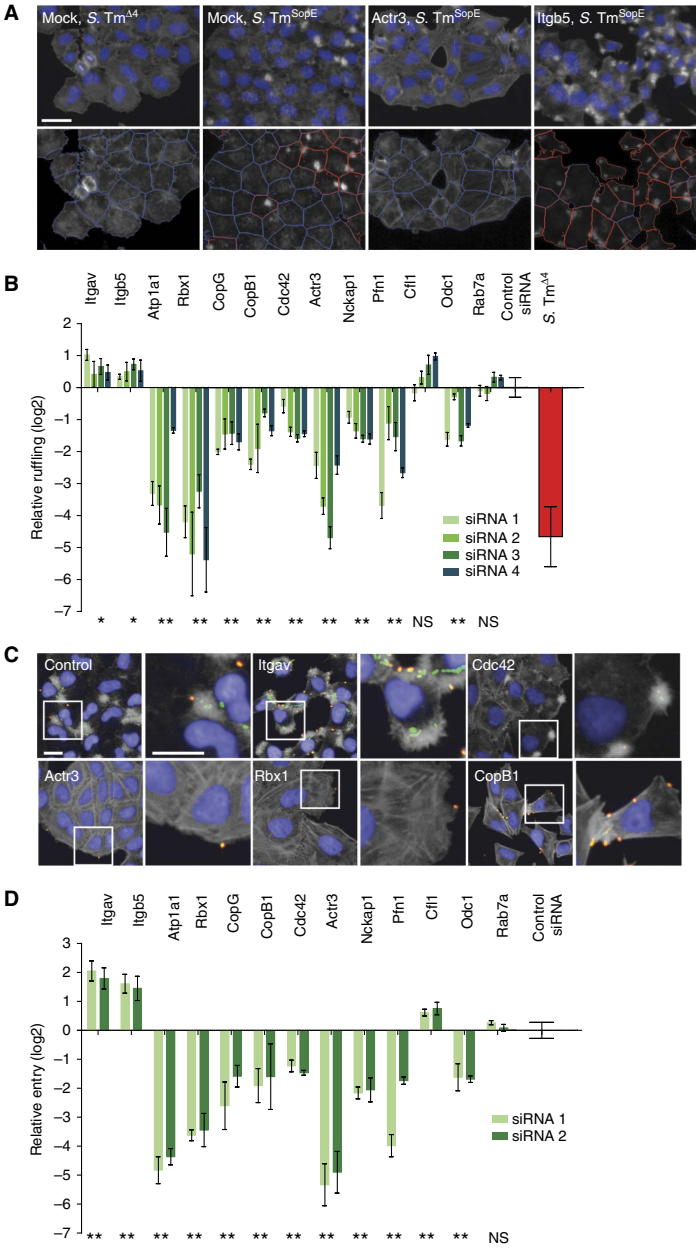


Figure 4 Depletion of Atp1a1 and Rbx1 results in strong inhibitory effects on ruffling and membrane closure. (**A**) Functionality of the image analysis of ruffling cells. Upper panels: details of the original automated microscopy images (blue=nuclei, gray=actin). Lower panels: results of the automated image analysis (red outlines=ruffling cells, blue outlines=non-ruffling cells; scale bar=20 μm). (**B**) Examples of the hits with a positive or a negative effect on ruffling induced upon *S.* Tm^{SopE} invasion. Multiple test correction for 298 genes was performed to obtain the *P*-value (**P*-value <0.1, ***P*-value <0.05, NS=not significant). (**C**) Fluorescence images showing cells treated either with control siRNAs or siRNAs directed against the indicated genes and infected with *S.* Tm^{SopE} (pM965) (red/green double stain=extracellular bacteria; blue=nuclei; gray=actin; green=intracellular bacteria; scale bar=20 μm). (**D**) Relative invasion of *S.* Tm^{SopE} into cells transfected with siRNAs directed against the indicated genes. Multiple test correction for 60 genes was performed on the results (***P*-value <0.05; NS=not significant).

endocytic vacuole independently of subsequent events in the eukaryotic cell. An automated quantification of intracellular bacteria was not possible using the images acquired with the available automated epifluorescence microscope (resolution: $\times 20$) due to the small size and spatial arrangements of cell-associated bacteria. We therefore opted for a manual quantification focusing on the 60 genes displaying the most pronounced invasion phenotypes (two siRNAs per gene). HeLa cells were infected with *S.* Tm^{SopE} (pM965), which constitutively express GFP. After 20 min, the cells were fixed and stained with DAPI (nuclei), TRITC-phalloidin (actin) and a *S.* Typhimurium LPS-specific antibody. Membrane closure was assessed by fluorescence microscopy, using differential antibody staining to distinguish extracellular bacteria (GFP^+, LPS^+) and internalized bacteria (GFP^+, LPS^-, Figure 4C).

Overall, the results obtained with this assay were similar to those obtained in the ruffling assay (compare Figure 4D and B; Supplementary Table SIII). This indicated that none of the tested genes were essential for membrane fusion or formation of the early endocytic vacuole. Furthermore, the vast majority of hits identified in our original invasion screen could be assigned to early steps of the infection process, i.e., binding, effector injection or ruffling, and did not seem to affect later maturation steps of the SCV. A notable exception was Rab7a. Rab7a depletion did not affect the membrane closure assay but had a significant phenotype in the original invasion assay. This is in line with its established role in maturation of the SCV (Meresse *et al*, 1999). However, besides Rab7a we were surprised to find only a few hits partially reducing SCV maturation, including the trafficking protein Vps39 and a subunit of the vacuolar ATPase (Atp6ap2, Supplementary Table SIII).

Step-by-step comparison identifies hits affecting multiple steps of the invasion process

With the large data sets available, we were able to follow each hit through the individual steps starting with binding, followed by effector injection, ruffling, membrane closure and SCV maturation, the readout of the modified gentamycin protection assay (Figure 5A). Depletion of ornithin decarboxylase 1 (Odc1; http://herkules.oulu.fi/isbn9514266315/) inhibited only the binding step. The same negative effect could be observed in all subsequent steps of the invasion process. Odc1 is therefore a bona fide binding hit. Similarly, Cdc42 meets the definition of a ruffling hit, as Cdc42 depletion did not affect binding or effector injection, but yielded similar levels of attenuation in the ruffling assay, the membrane closure assay and in SCV maturation. Rab7a represents a bona fide SCV maturation hit as the depletion influenced only the last step of the invasion process.

Importantly, many genes affected multiple steps, for instance binding and ruffling in the absence of subunits of the Arp2/3 (Actr3). Similarly, depletion of members of the COPI complex had a weak enhancing effect on binding and a specific negative effect on effector injection, and an additional negative effect on ruffling. Most likely, these effects were specific for COPI depletion, as all five subunits of the complex showed equivalent results (four oligos per gene; Supplementary Figure S2). The specific effect on injection is clearly illustrated by plotting the relative effect on binding efficiency against the relative effect on injection efficiency for each analyzed hit (Figure 5B). Thus, the COPI complex is an effector injection and a ruffling hit (Figure 5B; pink squares). Other examples of genes affecting more than one specific step of invasion include Pfn1, Actr3, Atp1a1 and Rbx1 (see Supplementary Table SIII; Supplementary Figure SIII).

Phenotypic clustering of the hits

To systematically analyze functional links between the 298 genes analyzed in detail (Supplementary Table SIII), we also used an automated clustering algorithm. We reasoned that (i) hits affecting the same cellular process should yield similar patterns of phenotypes in the step-specific assays. (ii) Hits affecting the same cellular process may interact either directly or indirectly. This could reveal cellular processes and molecular interactions, which may not have been noticed so far.

Cluster analysis of the normalized step-specific phenotypes yielded seven clusters (Figure 5C; Supplementary Table SV). Within each cluster, we automatically annotated the known functional interactions between the proteins/genes using the STRING 8.3 data base (Jensen *et al*, 2009).

The clusters c (Figure 5D) and d (Figure 5E) significantly enriched for known interactions, supporting the validity of our approach ($P<0.003$ versus random clusters of equal size). Cluster c, which harbors many binding and ruffling hits, included known regulators of membrane ruffling (Cdc42, NckAP1) as well as the ARF GTPase-activating factor Git2 (G-protein-coupled receptor kinase interactor 2; Cat-2), the proteasome, transcription factors (Myc, Max, Gtf2H1, Med4), several trafficking regulators (Sec13, Stx5, Stx17, Rab5c), proteins involved in cholesterol metabolism (Cyp27a1, TspO) and numerous genes not previously implicated in membrane ruffling (e.g., Tom1, CryZl1, Trim25). It will be of interest to determine how these factors may interact and whether they affect membrane ruffling directly or indirectly.

Cluster d harbors hits strongly affecting ruffling and vesicle maturation, including well-established actin regulators (ActR2, ActR3, Pfn1), as well as the NaK-ATPase (Atp1a1), Rbx1 and the COPI complex, which have recently been linked to regulation of the actin cytoskeleton (Valderrama *et al*, 2000; Wu *et al*, 2000; Barwe *et al*, 2005; Chen *et al*, 2009). However, their functional importance and possible interactions facilitating *S.* Tm invasion remain to be established. In conclusion, the cluster analysis revealed important information on established functional interactions (Figure 5D and E; lines connecting different hits) and possible novel interactions (no lines), thus providing a rich resource for future research on the host cell signaling modules driving each step of the invasion process.

The COPI complex affects the localization of cholesterol, gangliosid GM1 and the Rho GTPases Rac1 and Cdc42

Among our hits the COPI complex was unique by affecting effector injection and ruffling (Figure 5A). We therefore

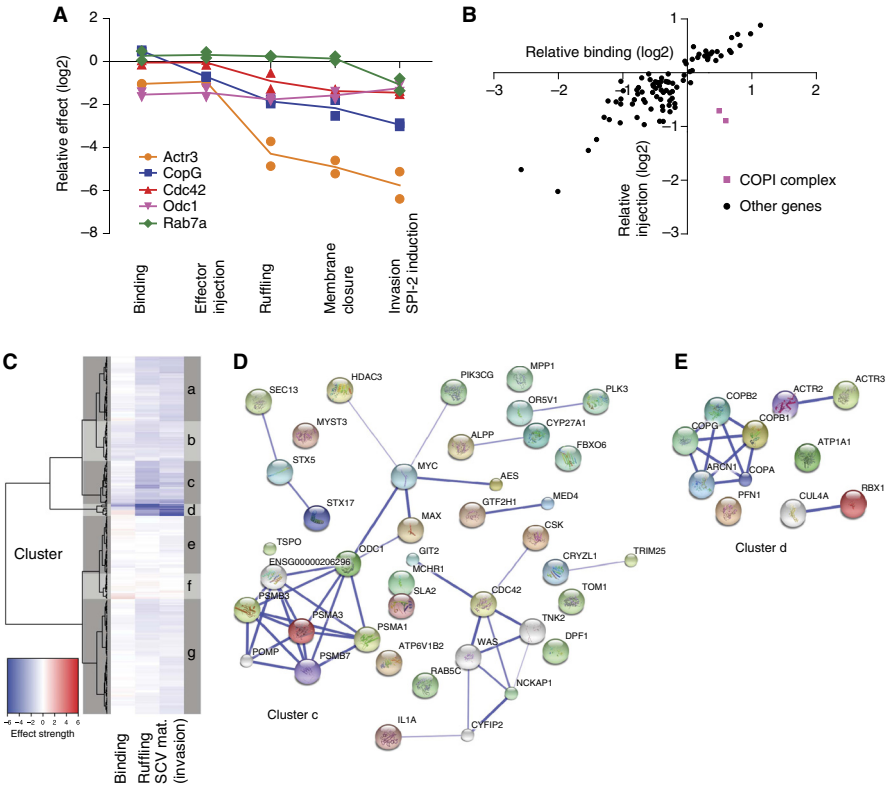


Figure 5 Classification of hits according to their profile in the different assays and cluster analysis of the results of the whole screen. (**A**) Overview showing the results of the different assays describing *Salmonella* invasion steps for selected genes. While Odc1 shows a pure binding phenotype, Cdc42 and Rab7a are examples for pure ruffling or maturation phenotypes. Several proteins show combined phenotypes: Actr3 depletion has effects on binding and ruffling, whereas depletion of CopB1 and CopG have inhibitory effects on effector translocation and ruffling. For consistency only the results of the two strongest siRNAs are shown. (**B**) Scatter plot of relative binding versus relative effector injection. A strong correlation between binding and effector injection is observed for most of the hits. The outliers CopB1 and CopG are stained in purple. Each point indicates the median of the two strongest siRNAs. The results for 90 genes are shown. (**C**) Heatmap and clustering dendrogram based on the 300 'candidate hits' obtained in the initial screen. (**D**, **E**) Functional interactions within the clusters were annotated using the STRING database (confidence cutoff=0.4, additional white nodes: five in cluster c and one in cluster d).

investigated the effects of the COPI complex depletion in detail. COPI is a central regulator of vesicular transport (Pucadyil and Schmid, 2009), suggesting that the depletion of the COPI complex might be explained by altered distribution of proteins important for *S*. Typhimurium invasion. Prime candidates for such proteins are the two Rho GTPases Cdc42 and Rac1 reported to have a role in SopE-mediated ruffling (Hardt *et al*, 1998; Friebel *et al*, 2001). Moreover, direct binding of Cdc42 to the γ subunit of the COPI complex has been observed (Wu *et al*, 2000). Redistribution of either Rho GTPase would explain the ruffling phenotype of COPI.

We therefore generated stable cell lines expressing Cdc42–GFP and Rac1–GFP, respectively. As shown in Figure 6A and Supplementary Figure S4, both Rho GTPases localized to the plasma membrane, to vesicles next to the nucleus and to smaller degrees to the cytoplasm. Quantification showed an enrichment of the GFP signal at the plasma membrane. Plasma membrane localization was even more pronounced for Rac1–GFP than for Cdc42–GFP (Figure 6A and B and Supplementary Figure S4A). After depletion of CopG or CopB1, both Rho GTPases were no longer found enriched at the plasma membrane, but accumulated instead in an intracellular compartment (Figure 6A and B, Supplementary Figures S4 and S7). The latter is probably equivalent to the prominent vesicular compartment accumulating markers for the trans-Golgi network (TGN), the endoplasmatic reticulum–Golgi intermediate compartment (ERGIC), the Golgi apparatus and recycling endosomes, but not for early or late endosomes or lysosomes, which had been identified recently in β-Cop (CopB1)-depleted cells (Styers *et al*, 2008). Depletion of several other proteins including Rbx1, Atp1A1, Actr3 did not change the localization of the Rho GTPases (data not shown), making this Rho GTPase relocalization phenotype unique among the 'ruffling' hits.

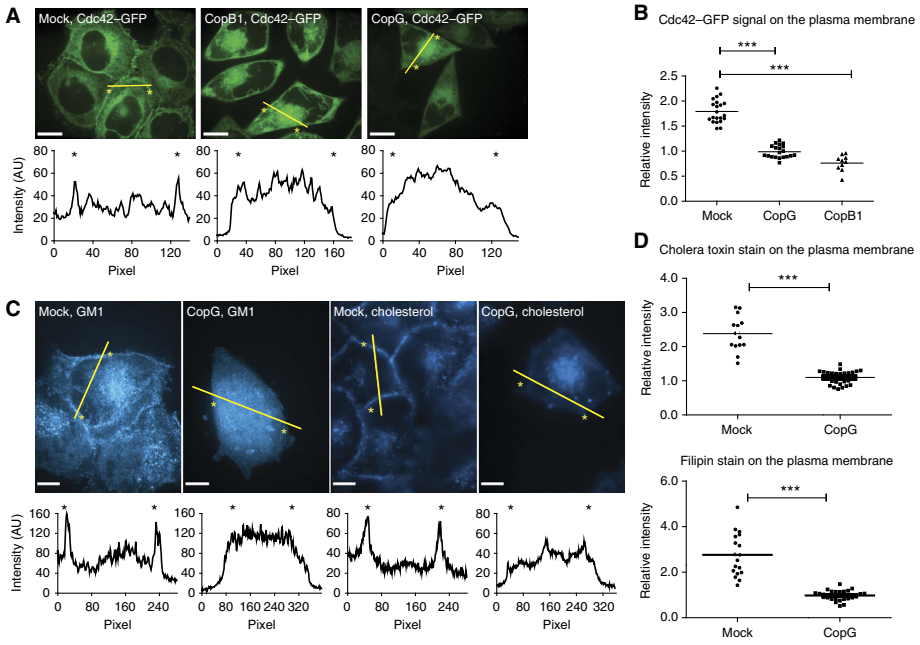


Figure 6 Cdc42, sphingolipid GM1 and cholesterol are mislocalized after depletion of the COPI complex. (**A**) Confocal images showing the localization of Cdc42–GFP in water-transfected cells (left), cells transfected with siRNA directed against *copB1* (middle) or *copG* (right). The intensity plots along the lines indicated in yellow are shown below the pictures; the position of the membrane is indicated by an asterisk; scale bar=10 μm. (**B**) Quantification of the Cdc42–GFP signal on the membrane relative to the cytosol (***$P<0.005$). (**C**) Confocal images showing the distribution of sphingolipid GM1 (left two panels) and cholesterol (right two panels) in the cells. For each staining, water-transfected cells are shown on the left side and cells transfected with siRNA against *copG* on the right side; scale bar=10 μm. (**D**) Quantification of the relative cholera toxin β/filipin staining signal at the membrane compared with the cytosol (***$P<0.005$).

Next, it was tested whether altered lipid composition of the plasma membrane might explain Rho GTPase relocation upon COPI depletion. Cholesterol-enriched membrane domains are known to bind activated Rho GTPases via their geranylgeranyl anchors (Palazzo *et al*, 2004; del Pozo *et al*, 2004) and COPI vesicles are implicated in membrane sorting of sphingolipids and cholesterol. Membrane components such as the sphingolipid GM1 and cholesterol can be visualized by staining with the β-subunit of cholera toxin or filipin, respectively. Under normal conditions, both markers mainly localized to the plasma membrane and to some intracellular vesicles. In contrast, after depletion of CopG, the cell membrane was no longer stained by either of the dyes, but the signal accumulated in a large intracellular compartment, colocalizing with Cdc42 and Rac1 (Figure 6C and D, Supplementary Figures S5 and S6). These results suggest that cholesterol and GM1 are retained in an atypical compartment formed after depletion of different subunits of the COPI complex. This would lead to a mislocalization of Rac1 and Cdc42, effectively depleting both proteins from the cell membrane. In agreement with these data, the size of the ruffles formed at the site of *Salmonella* binding was significantly decreased after depletion of CopG (Figure 7A).

To demonstrate that the injection and/or the ruffling defects were attributable to altered membrane lipid composition, we replenished cholesterol in COPI-depleted cells by adding cholesterol complexed to methyl-β-cyclodextrin (MβCD) 5 h prior to the infection of the cells. We infected the cells and non-complemented controls for 10 min with *S.* Tm^{SopE} and measured the diameter of the ruffles formed by a single bacterium. Indeed, replenishing led to a significant increase in the size of the ruffle compared with the CopG-depleted cells not replenished with cholesterol (Figure 7B).

The central role of COPI in cholesterol and Rho GTPase localization was also confirmed by inhibitor experiments. Thereby, pretreatment of cells with either geranylgeranyltransferase inhibitor (GGTI, inhibits geranygeranylation of the Rho GTPases thus dislocalizing them from the plasma membrane) or MβCD (depletes cholesterol from the plasma membrane) decreased *S.* Typhimurium invasion (data not shown). Importantly, both inhibitors could not further diminish *S.* Tm invasion in CopG-depleted cells, suggesting that both cholesterol and Rho GTPases were already missing at the plasma membrane due to COPI inactivation. This confirmed that the ruffling defect of CopG-depleted cells was indeed attributable to the mislocalization of the Rho GTPases and cholesterol. Further work would be required to determine the exact molecular mechanism explaining this defect, e.g., defective Rho GTPase membrane localization in resting cells or reduced Rho GTPase recruitment after pathogen binding. Taken together, our results

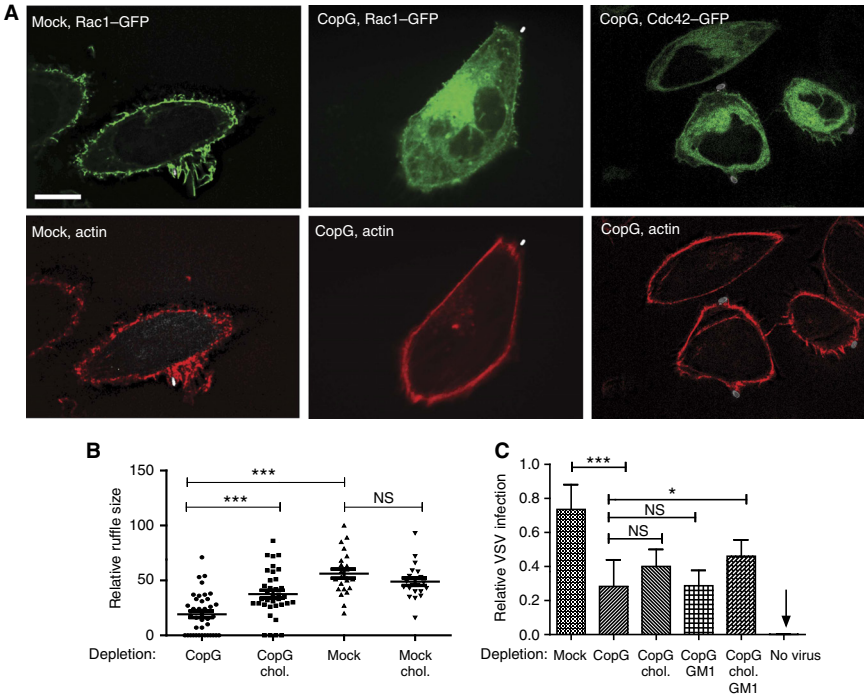


Figure 7 CopG depletion abolishes actin recruitment to the site of *Salmonella* binding and ruffling. (**A**) Confocal images showing the distribution of actin and Rho GTPases upon *S.* Tm^{SopE} invasion for 10 min in stable Rac1- or Cdc42-GFP-expressing cell lines. The cells were transfected either with water (mock) or with siRNA directed against *copG*; scale bar=10 μm. (**B**) Size of membrane ruffles after infection with *S.* Tm^{SopE} for 10 min of CopG-depleted cells with or without cholesterol replenishing. The ruffle size was measured in the focal plane showing the largest diameter of the ruffle. (**C**) Infection of cells either transfected with water (mock) or with siRNA directed against *copG* by VSV for 5 h. The relative invasion was quantified using a virus construct, which leads to a green fluorescent staining of the cytoplasm upon invasion. Quantification of the amount of infected cells in untransfected cells (mock), or cells depleted of CopG, which were either complemented with cholesterol, with GM1 or with both prior to infection are shown (***$P<0.005$; *$P<0.05$; NS=not significant).

suggest a novel role for the COPI complex in the localization of cholesterol and Rho GTPases at the plasma membrane.

COPI-dependent membrane lipid composition also affects vesicular stomatitis virus infection

Strikingly, a large number of bacterial and viral pathogens share two phenotypes observed for SopE-mediated host cell invasion, i.e., the requirement for cholesterol-enriched microdomains in the host cell membrane and reduced infection efficiency upon COPI depletion. This includes *Staphylococcus aureus* (Ramet *et al*, 2002; Potrich *et al*, 2009), *Escherichia coli* (Ramet *et al*, 2002; Philips *et al*, 2005; Riff *et al*, 2005), *Listeria monocytogenes* (Seveau *et al*, 2004; Agaisse *et al*, 2005; Cheng *et al*, 2005; Gekara *et al*, 2005), *Mycobacterium fortuitum* (Philips *et al*, 2005), *Mycobacterium tuberculosis* (Munoz *et al*, 2009), *Chlamydia caviae* (Derre *et al*, 2007), *Chlamydia trachomatis* (Jutras *et al*, 2003; Elwell *et al*, 2008), influenza virus (Hao *et al*, 2008; Konig *et al*, 2010) and hepatitis c virus (Tai *et al*, 2009; Popescu and Dubuisson, 2010). However, a functional link between both phenotypes had never been established.

Our results suggest that the dependence on both COPI and cholesterol membrane microdomains might be functionally linked. To test this hypothesis, we analyzed the effect of COPI depletion and of replenishing cholesterol and sphingolipids such as GM1 on host cell infection by the vesicular stomatitis virus (VSV). VSV is an enveloped, single-stranded, negative-sense RNA virus of the family Rhabdoviridae. It infects a wide range of host cells and the existence of a specific receptor is still not entirely clear (Schlegel *et al*, 1983; Coil and Miller, 2004). Host cell entry occurs via clathrin-dependent endocytosis, and low-pH-mediated alterations in the virus glycoprotein lead to membrane fusion (Sun *et al*, 2005).

Indeed, CopG depletion significantly interfered with VSV host cell invasion (Figure 7C). Moreover, the infection efficiency could partially be restored by replenishing GM1; this effect was even more pronounced after replenishing both, GM1 and cholesterol. In contrast, replenishing just cholesterol without GM1 was insufficient for rescuing VSV invasion into CopG-depleted cells. Thus, some differences exist between the host cell membrane lipid requirements for VSV and SopE-dependent *S.* Typhimurium infection. Nevertheless, these data

indicate that COPI-dependent control of the membrane lipid composition is of importance for vastly different pathogens.

Discussion

The genome-scale siRNA screen identified host cell factors mediating SopE-dependent *S.* Typhimurium invasion into epithelial cells. Step-specific follow-up assays detected at least one important host factor for each step, except for membrane closure. This step-specific analysis allowed the functional classification of the respective host cell factors, confirming well-established ones, revealing new ones and suggesting novel functional links between them.

Many hits identified in our screen affected the binding step. This was unexpected and suggested that binding was rate-limiting under our experimental conditions. Among the binding hits, we found numerous known regulators of actin polymerization, including Pfn1, Cofilin 1 or the Arp2/3 complex, as well as other host cell proteins implicated in cell shape or actin regulation (e.g., Itgb5, Itgav; Supplementary Tables SIII and SIV). The actin cytoskeleton determines cellular shape, membrane stiffness and the presentation of surface proteins, three parameters likely to affect pathogen binding. Therefore, genes identified as binding hits in the screen might affect these three cellular characteristics. It will be of interest to determine the underlying molecular mechanisms.

The ruffling step was affected by numerous hits and allowed us to refine the model of this key step of host cell invasion (Figure 8). In the host cell, SopE activates Cdc42 and Rac1. The presented data suggest that the Rho GTPases must be localized at the plasma membrane in order to trigger ruffling. Together with Nap1, Sra1 and Pir121 (reviewed in Schlumberger and Hardt, 2006) they lead to activation of N-Wasp or WAVE complex and subsequently to Arp2/3 complex-mediated actin polymerization. Our data suggest that this is further enhanced by Pfn1, a well-known factor delivering actin monomers to sites of actin polymerization, but not previously implicated in *Salmonella* infection (Pollard and Cooper, 2009). The actin polymerization step is negatively regulated by actin-depolymerizing proteins such as cofilin 1 (Dai *et al*, 2004; McGhie *et al*, 2004). Furthermore, it has been observed that Cap1 is an additional factor stimulating the breakdown of actin fibers, thus negatively regulating *Salmonella* invasion. This supports that the induction of membrane ruffling is the result of a complex cascade of actin polymerizing and depolymerizing events.

In most cases, the ruffling defect led to a corresponding defect in host cell invasion. However, in a few cases (e.g., proteasome complex, Rbx1), the ruffling defect was much more pronounced than the invasion defect. This might be explained by the existence of two independent modes of host cell invasion by *S.* Typhimurium. Indeed, a ruffling-independent mode of *S.* Typhimurium invasion has recently been described (Hanisch *et al*, 2010). The former group of ruffling hits might affect both modes of host cell invasion. In contrast, the proteasome and Rbx1 might have specifically affected the ruffling-mediated invasion, leaving the ruffling-independent invasion process untouched. This will be an interesting topic for future studies.

The COPI complex also affected membrane ruffling which was at least partially attributable to the reduced cholesterol/sphingolipid content of the plasma membrane upon CopB or CopG depletion. Depletion of COPI components resulted in cholesterol and GM1 relocalization to a large intracellular structure, probably representing the 'common compartment' observed in a recent study upon β-Cop-depletion harboring TGN, ERGIC, Golgi and recycling endosome membrane protein markers (Styers *et al*, 2008). To the best of our knowledge, the drastic effect of COPI depletion on cholesterol trafficking has not been reported previously. Nevertheless, the COPI complex has been implicated in other trafficking events. Its role for the retrograde transport from the Golgi to the endoplasmic reticulum is best understood (Bethune *et al*, 2006; Beck *et al*, 2009). However, the reason for cholesterol relocalization in CopG-depleted cells is still unclear. On the one hand, cholesterol is synthesized in the smooth endoplasmic reticulum and has a complex trafficking

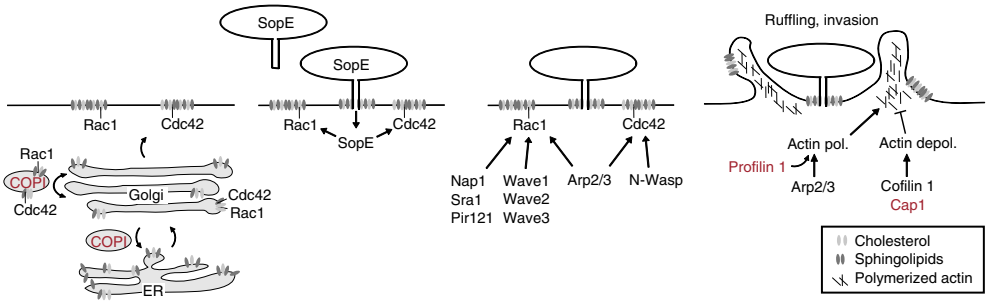


Figure 8 Model for *S.* Typhimurium entry into HeLa cells. In order to invade, *S.* Typhimurium binds onto cells and injects a cocktail of effectors. This study investigates the action of the key effector SopE. SopE mediates the GTP nucleotide exchange and thereby the activation of the Rho GTPases Rac1 and Cdc42. This leads to the activation of the Arp2/3 complex, actin polymerization cellular ruffling and bacterial entry. The screen identified key regulators of this process, including Nap1, Cdc42, Profilin 1 and the Arp2/3 complex, as well as proteins mediating actin depolymerization, a process probably counteracting the activity of Rac1 and Cdc42. The correct localization seems to be as important as the activation of the Rho GTPases. Upon depletion of the COPI complex, cholesterol and Rho GTPases are mislocalized away from the cell membrane, leading to a much less efficient *S.* Typhimurium entry. Proteins newly identified to be implicated in *S.* Typhimurium entry are indicated in red. Proteins that have already been described to have a role in the infection and were found in the screen are indicated in black.

route, involving vesicular and non-vesicular trafficking steps (Ikonen, 2008). A deficient transport/sorting from the Golgi apparatus to other compartments might thus explain the lack of cholesterol and sphingolipids on the cell membrane and its accumulation on the common compartment. On the other hand, cholesterol is endocytosed at the cell membrane and traffics to endosomes. Subsequently, certain lipids must leave the path to the lysosome, as the cholesterol content of lysosomal membranes is low. By trapping the recycling endosomes in the common compartment, CopG depletion might interfere with cholesterol resorting to the plasma membrane. Defining the underlying mechanism of cholesterol trafficking will be an important task for future work.

Upon CopG or CopB depletion, the Rho GTPases Cdc42 and Rac1 were mislocalized in a similar manner as cholesterol and GM1. These Rho GTPases are geranylgeranylated and this modification localizes them to cholesterol-enriched lipid microdomains (del Pozo *et al*, 2004; Palazzo *et al*, 2004). In line with this hypothesis, disrupting geranylgeranylation of Rho GTPases with a chemical inhibitor released Rac1–GFP and Cdc42–GFP from the plasma membrane and prevented relocalization to the 'common compartment' in COPI-depleted cells (data not shown). On the basis of these observations, it is tempting to speculate that cholesterol/sphingolipid relocalization might result in the mislocalization of additional proteins, i.e., of cholesterol/sphingolipid binding membrane proteins such as Rho GTPases in COPI-depleted cells. Possibly, the cholesterol/sphingolipid relocalization phenotype might help to decipher proteins transporting these membrane lipids from proteins that are transported passively by them.

The reduced effector protein injection into COPI-depleted cells might be directly attributable to reduced cholesterol levels in the plasma membrane. The *S.* Typhimurium protein SipB, located at the tip of the translocon, is known to bind cholesterol with high affinity *in vitro* (Hayward *et al*, 2005), suggesting that cholesterol may directly activate bacterial effector injection by T1. In line with this hypothesis, artificial liposomes containing cholesterol and sphingolipids were shown to activate the Type III secretion system of *Shigella flexneri*, a closely related enteropathogenic bacterium (van der Goot *et al*, 2004). In addition, the *Shigella* effector IpaB, a homolog of SipB, could bind CD44 in cholesterol-enriched microdomains (Lafont *et al*, 2002), pointing to a general role of host cell membrane lipids in activating Type III secretion systems. Thus, the mislocalization of Rho GTPases, i.e., Cdc42, together with reduced effector injection efficiency into CopG-depleted cells provides satisfactory explanation for the defect in *S.* Tm^{SopE}-induced ruffling.

In this study, five assays were used to quantify effects on particular steps of *S.* Typhimurium invasion into HeLa cells. Four of these assays could be performed in a high-throughput format. A detailed discussion of the specificities of the assays is available upon request. Owing to the combination of gentamycin protection and intracellular GFP expression, the modified gentamycin protection assay (Figures 1 and 2) has a high specificity and a very robust performance, thus enabling genome-scale experiments. The binding or the ruffle assays are less robust, and require careful controls excluding the presence of staining artifacts, which may yield 'false positives' (i.e., over-estimation of the number of affected cells). The analysis of a large number of cells contributed to further enhancing the robustness of these assays. In the case of the ruffle assay, false positive might have occurred if the siRNA affected ruffle-like changes in the actin cytoskeleton, even in the absence of bacteria. However, the identification of step-specific genes and the high concordance of all other genes in the different subsequent assays (Supplementary Figure S2) argues that all assays were reasonably robust.

Contrary to our expectations (Schlumberger *et al*, 2006), the Rho GTPase Rac1 and N-WASP, linking Cdc42 and the Arp2/3 complex, have not been discovered by our initial genome-scale screen. Most likely, they belong to the 'false negatives', which are commonly encountered in genome-scale screens. When four new and partially different siRNAs directed against each gene were reordered, an inhibitory effect on *S.* Tm invasion upon depletion could be demonstrated (data not shown). Thus, due to an improved siRNA design and/or altered experimental conditions in later experiments, we could correct these false-negative genes (data not shown).

In conclusion, the presented genome-scale screen of SopE-mediated invasion revealed an array of hits enhancing the current model of triggered host cell invasion and a set of proteins not associated with *S.* Tm invasion so far. In addition, follow-up assays on *S.* Tm and VSV and published work implicate the COPI-mediated control of cholesterol and sphingolipid distribution to the host cell plasma membrane as a common mechanism affecting infection by bacterial and viral pathogens. Such host cellular functions required by phylogenetically diverse groups of pathogens are of great interest for understanding the evolution of infectious disease. Moreover, they are of practical interest, as they might offer starting points for developing novel anti-infective therapeutics with a very broad range of biological activity.

Materials and methods

Bacterial strains and plasmids

All *S.* Typhimurium strains used were isogenic derivatives of SL1344 (SB300) of *S. enterica* subspecies I serovar Typhimurium (Supplementary Table SI; Hoiseth and Stocker, 1981). Strains M701 (Muller *et al*, 2009), M566 (Ewen *et al*, 1997), SB161 (Kaniga *et al*, 1994) and M1114 (Schlumberger *et al*, 2007) have been described previously. M1128 was constructed by P22-phage transduction of the *sipA*::*sipA-M45-tem1* allele of strain 1104 into strain M713. M1104 has been described previously (Schlumberger *et al*, 2007) and strain M713 was constructed by p22-phage transduction of the *sseD*::*aphT* allele from MvP101 (*S.* Typhimurium ATCC 14028 derivate; Medina *et al*, 1999) into strain M712 (Ehrbar *et al*, 2004). Plasmids pM965 (Stecher *et al*, 2004), pM975 (Hapfelmeier *et al*, 2005) and pEGFP-C3/Rac1WT (Hage *et al*, 2009) have been described previously. Plasmid pEGFP-C1/Cdc42hsWT was kindly provided by Dr Klaudia Giehl.

For the infection of HeLa cells, bacteria were grown in LB broth supplemented with 0.3 M NaCl and 50 mg/l Streptomycin (AppliChem) for 12 h at 37 °C and subcultured for 4 h.

Cell cultures

HeLa cells were grown in DMEM (PAA Laboratories) supplemented with 10% FCS (Omnilab) and 50 mg/l Streptomycin (AppliChem) at 37 °C and 5% CO_2. HeLa cells were stably transfected with pEGFP-C3/Rac1WT or pEGFP-C1/Cdc42hsWT and clones were

maintained in medium supplemented with 500 mg/l Neomycin (AppliChem).

siRNA transfection

For siRNAs a reverse transfection protocol was used. In 96-well plates (μ-clear bottom, Greiner Bio One), 2 μl of 1 μM siRNA was added to 8 μl Opti-MEM (Invitrogen) yielding a final siRNA concentration of 20 nM (after addition of cells). Lipofectamine 2000 (Invitrogen) was diluted 1:200 in Opti-MEM and incubated for 15 min at room temperature. A quantity of 10 μl per well were added and incubated for another 15 min at room temperature. These plates, henceforth referred to as cell plates, were either directly used or frozen at −80 °C. HeLa cells were seeded using 1800–2000 cells in 80 μl per well, followed by an incubation of 3 days at 37 °C and 5% CO_2. For half-size plates (μ-clear bottom, half area, Greiner Bio One), all numbers were reduced to 60%. For the genome-scale screen, 384-well plates (μ-clear bottom, Greiner Bio One) were used. The volume of the different reagents was reduced to 50%. The final siRNA concentration was changed to 50 nM and Lipofectamine was used at a concentration of 1:100.

siRNA libraries

SiRNAs were ordered from Qiagen. For the genome-scale screen the 'druggable genome' library (Version 2.0) comprising 6978 genes and 3 siRNAs per gene were used. Data were normalized platewise to the median of all siRNAs on the respective plate. Experiments were performed in triplicates and the median for each siRNA calculated, either with or without platewise *z*-score correction. The median value for the three siRNAs per gene was defined as the final read-out for each gene. A hit was defined as 1.5 times the interquartile range from the median of the entire data set.

For a confirmatory screen and secondary assays, another library was assembled based on hits from the *z*-score corrected list (182 genes), with 80 additional hits from the uncorrected list as well as 38 additional genes with high biological probability according to our screen results, but not present in the druggable genome library. Altogether, this secondary library contained 298 genes and 4 siRNAs per gene. To enable comparison of our data with the large screen, six genes with the smallest deviation of all three siRNAs from the median of the whole screen were selected as controls for the secondary library. In total, 12 control siRNAs were used: Hs_RAB30_8, Hs_RAB30_7, Hs_P53AIP1_5, Hs_P53AIP1_6, Hs_DNAJC6_1, Hs_DNAJC6_5, Hs_ARNT2_1, Hs_ARNT2_3, Hs_FBXL2_1, Hs_FBXL2_2, Hs_ANGPTL3_2 and Hs_ANGPTL3_1. If not otherwise indicated, the median of these 12 siRNAs was used for platewise normalization of secondary assays. In all siRNA experiments, siRNAs against subunits of the Arp complex (Arc21, ArpC3_3, ArpC3_5), CDC42 (CDC42_7, CDC42_10) and CFL1 (CFL1_3, CFL1_5, CFL2_5) were used as additional controls of siRNA effects on *S.* Tm infection. In addition, siRNAs with known cytotoxic effects were used to control for siRNA-transfection efficiency (PLK1_2 and EG5).

Immunoblot analysis

HeLa Kyoto cells were seeded in 24-well plates and transfected as described above, using a final siRNA concentration of 20 nM. Cells of two wells transfected with siRNAs against the same proteins were lysed in 50 μl Laemmli sample buffer and incubated for 10 min at 95 °C. Whole samples were subjected to 6–16% SDS–polyacrylamide gel electrophoresis. In order to confirm the depletion efficiency, the following antibodies were used: anti-actin (Santa Cruz sc-1615 and Sigma A3853), anti-calnexin (Santa Cruz sc-6465), anti-Cdc42 (BD Laboratories 610929), anti-Cfl1 (Abcam ab11062), anti-CopB1 (Abcam ab6323), anti-CopG (Gentex GTX105331), anti-Itgav (Santa Cruz sc-9969), anti-Odc1 (Abcam ab50269), anti-Pfn1 (Abcam ab50667), anti-Rab7a (Abcam ab50533) and anti-Rbx1 (anti-Roc1, Abcam ab2977). The antibody against Atp1a1 was kindly provided by Dr Jack Kaplan. Anti-Actr3 and anti-Nckap1 (Nap1) were kind gifts from Dr Klemens Rottner. Proteins were detected using either goat anti-rabbit immunoglobulin G-horseradish peroxidase (Jackson 111-035-003) or goat anti-mouse immunoglobulin G-horseradish conjugates (Sigma A4416) and using ECL substrate (Amersham), as recommended by the manufacturer. Densitometric analysis on blots was performed with the QuantityOne software (Bio-Rad). The density in siRNA-treated groups was normalized to the density of their actin or calnexin levels and to corresponding control siRNA-treated groups.

Modified gentamycin protection assay

HeLa cells were infected with *S.* Tm^{SopE} using an m.o.i of 64 for the genome-scale and 32 for the confirmatory screen. Bacteria were either added using a 384-needle head of a robot (Tecan) or a multi-channel pipette. After 22 min, the medium was replaced with medium containing gentamycin (400 μg/ml; AppliChem) followed by 4 h of incubation at 37°C, fixation using 4% paraformaldehyde (PFA) and staining with DAPI (10 μg/ml) in 0.1% Triton X-100. For the computation of *S.* Tm infection, an assay using automated microscopy and automated images analysis was employed: nine images per well were acquired using a cellWoRx microscope (Applied Precision) and a × 10 objective. Images were analyzed using the open source program CellProfiler (Carpenter *et al*, 2006) as well as customized algorithms. In brief, nuclei were analyzed using the IdentifyPrimary module of CellProfiler. Nuclei were expanded with the CellProfiler-module IdentifySecondaryObjects. Spots were identified with the IdentifyPrimary module using a fixed threshold for the intensity in the GFP channel. Finally, cells and spots were superimposed and cells containing at least one spot were counted as infected. In a preprocessing step, the optimal threshold in the GFP channel separating signal and noise best was automatically calculated by comparing gentamycin pretreated wells and mock-treated (water instead of siRNA) wells.

Binding assay

The binding assay has been described recently (Misselwitz *et al*, 2011). Cells were pretreated with siRNAs as described for the confirmatory screen (Figure 2). In brief, cells were infected with *S.* Tm^{A4} (m.o.i. of 82) for 6 min using a multi-channel pipette followed by three washing steps in DMEM containing 10% FCS and PFA fixation using a microplate dispenser (WellMate, Thermo Scientific). Subsequently, bacteria were stained by indirect immunofluorescence using an anti-LPS antibody and a FITC-coated secondary antibody. Nuclei were stained using DAPI. Images were acquired on an Image Express microscope (Molecular devices) using a × 4 objective. Image analysis followed similar algorithms as described for the modified gentamycin protection assay. In a second analysis step, the Enhanced CellClassifier (Misselwitz *et al*, 2010) was used for recognition of mitotic nuclei and neighbors of mitotic nuclei. For the final analysis, nuclei were thus purged from mitotic nuclei and their neighbors.

Ruffling assay

Cells were infected with *S.* Tm^{SopE} (m.o.i. 250) for 6 min. Subsequently, cells were fixed with 4% PFA and stained with DAPI and TRITC-Phalloidin (0.5 μg/ml; Sigma). Images were subsequently acquired on a BD Pathway microscope using a × 20 objective in the DAPI and the TRITC channel. For image analysis, nuclei and cells were recognized as described for binding, except that for the definition of cell borders the information of the actin channel was also used. Subsequently, intensity and texture of cells and nuclei were measured using predefined CellProfiler modules. In addition, customized modules were designed detecting small areas with bright intensity in the actin channel. Details of our customized features and the analysis pipeline used are available upon request. After image analysis, data were imported in the program Enhanced CellClassifier (Misselwitz *et al*, 2010), which facilitates usage of a machine learning algorithm (SVM with a radial basis function). Thereby, the user first needs to label cells as either ruffling or non-ruffling. These labels as well as measurements for the corresponding cells are recorded by the program. Later each imaged cell can be assigned by the trained model to a particular class (i.e., ruffling or non-ruffling). It should be noted

that ruffle intensity cannot be scored by the Enhanced CellClassifier. For each of the three replica of the screen, more than 10 000 cells were trained as ruffling or non-ruffling. Mitotic cells were consistently trained as non-ruffling. During training, we focused on correct recognition of non-ruffling cells. After optimization of the critical parameters C and gamma of the SVM algorithm, a fivefold cross-validation accuracy of 94% was achieved.

Membrane closure assay using differential outside staining

HeLa cells were infected with *S.* Tm^{SopE} carrying plasmid pM965 for constitutive GFP expression (m.o.i. of 16–33) for 20 min and immediately fixed after infection. Bacteria were stained by indirect immunofluorescence without permeabilization using an anti-LPS antibody and an FITC-coated secondary antibody. Nuclei and actin were visualized using DAPI and TRITC Phalloidin after permeabilization. For each well, cell-associated extracellular and intracellular bacteria were quantified for 100–200 nuclei. Subsequently, the average number of invaded *S.* Tm^{SopE} per cell was calculated. For normalization, 7 of the 12 control siRNAs were evaluated for each plate. *Salmonella* invasion was thus quantified for the strongest two siRNAs of our 60 top hits by two independent observers. Key hits shown in the diagrams were quantified in three independent experiments.

Effector injection assay

HeLa cells were infected with *S.* $Tm^{SipA\text{–}TEM}$ (m.o.i. 125) for 6 min, followed by two washing steps in Hank's buffered salt solution (HBSS) containing gentamycin and chloramphenicol and 30 min incubation at room temperature. A CCF2 loading kit was purchased from Invitrogen and CCF2-AM was dissolved in DMSO at a 1 mM concentration and stored in aliquots at −80°C. Immediately before the assay, substrate loading solution was prepared by mixing the CCF2-AM stock solution with proprietary solutions B and C (Invitrogen) at a ratio of 1:10:156. Plates were emptied and 50 µl HBSS without antibiotics was added to the cells, followed by the addition of 10 µl substrate loading solution. Luminescence was quantified 120 min after addition of the loading kit using a Victor3 Multilabel Plate Reader (Perkin Elmer). The ratio of the blue (450 nm) and the green signal (520 nm) was calculated after platewise background correction using the average of three wells without cells.

Post-processing analysis of data

GFP invasion, ruffling and binding experiments were performed in three independent experiments. The effector injection assay was carried out in triplicates in two independent experiments; therefore, the data summarize six data points. In all experiments, siRNA controls were also conducted as well as specific controls for the various assays used. As effects mediated by *S.* Typhimurium (i.e., binding, ruffling, effector injection) were stable for a wide range of cell densities (data not shown), a correction for the number of nuclei or siRNA toxicity was not necessary. Statistical analysis was carried out using the Matlab implementation of the Mann–Whitney *U*-test, comparing the values for each siRNA tested with the control siRNAs. Multiple test correction was performed using the Matlab implementation of the Benjamini and Hochberg (1995) method.

Fluorescence staining and image acquisition for high-resolution microscopy

HeLa cells were seeded on coverslips 1 day before infection and then fixed with 4% PFA (Sigma) for 15 min and permeabilized with 0.1% Triton X-100 (Sigma) for 5 min. Cells were blocked with 3% BSA and incubated with the appropriate antibodies (TRITC Phalloidin). The coverslips were then mounted with Vectashield (Reactolab, SA) and imaged with a × 100 objective using a Zeiss Axiovert 200 m inverted microscope with an Ultraview confocal head (Perkin Elmer), a krypton argon laser (643-RYB-A01, Melles, Griot, Didam, The Netherlands) and a PLAN-Apochromat × 100 oil objective with an aperture setting of 1.3 (Zeiss). Infrared, red and green fluorescence was recorded confocally, whereas blue fluorescence (DAPI and filipin) was recorded by epifluorescence microscopy. Images were then deconvolved with Volocity 5.2.0 (Improvision, Conventry, UK) for 25 iterations.

Quantification of membrane localization of different proteins

Lines were randomly drawn through images of the different channels, and the intensity along these lines was measured by using the Intensity measurement tool of ImageJ. The data were analyzed using a MatLab program with the following algorithm: On each line the first local maximum was identified. This first maximum corresponds to the cell membrane intensity. Next, to investigate the intracellular intensity, an average over 75 pixels was calculated. To ensure a sufficient distance from the cell membrane, the starting point was defined as the first local minimum for which both adjacent local minima had a higher intensity. Finally, the ratio of both intensities was compared. All maxima and minima were visualized on the graphs and manually verified.

Treatment of cells with MβCD, GGTI and replenishment of cholesterol and GM1

Cells were seeded in DMEM supplemented with 10% FCS at an appropriate density (13 000 HeLa cells well of a 24-well plate or 2000 cells per well of a 96-well plate), subjected to siRNA treatment (20 nM siRNA, Lipofectamine in a dilution 1:200 in Opti-MEM medium) and pretreated for 24 h with GGTI at a concentration of 10 µM or with MβCD at 10 mM for 1 h. For the replenishing experiments, cells were incubated for 17 h with 3.5–14 µg/ml GM1 complexed to fatty acid-free BSA and/or for at least 5 h with 15 µg/ml cholesterol complexed to 0.37 mM MβCD.

Quantification of the ruffle size

Depletion of CopG and cholesterol replenishment was performed as described above, and cells were infected with *S.* Tm^{SopE} at an m.o.i. of 60 for 10 min. Only ruffles containing a single intra- or extracellular bacterium were considered. Cells were then prepared for immunofluorescence analysis as described above. Automated identification and measurements of ruffles are challenging due to a cellular variability of the signal of the actin channel, for instance, in different stages of the cell cycle. We therefore opted for a manual quantification. Stacks were acquired for all ruffles and the largest diameter of each ruffle was measured in the xy-plane using ImageJ. Future developments allowing more advanced image analysis techniques might allow for automated and unbiased ruffle quantification for large numbers of cells.

Infection with VSV

Depletion of CopG and cholesterol and GM1 replenishment was performed as described above. The cells were infected with a VSV strain overexpressing GFP (Pelkmans *et al*, 2005) for 5 h, fixed and analyzed by automated microscopy. Images were acquired on automated wide-field cellWoRx microscopes (Applied Precision) with a × 10 objective. Per well, 5 × 5 directly adjacent images were taken for all virus infection assays, covering over 85% of each well surface. Image-based auto-focusing was performed on the DAPI signal for each imaged site. The images were analyzed with CellProfiler combined with a software published previously for supervised classification of cellular phenotypes (Ramo *et al*, 2009), and image analysis methods published elsewhere (Snijder *et al*, 2009). The CellProfiler image analysis pipeline was used as follows: first, nuclei objects were identified based on the DAPI stain. Next, cell boundaries were estimated using nuclear expansion. Standard CellProfiler texture, intensity, size and shape features were extracted from nucleus and cell regions, as well as from complete images, for both the DAPI and virus GFP signal. We next applied supervised machine learning using the open source SVM learning tool CellClassifier (Ramo *et al*, 2009), to identify virus-infected cells.

Phenotypic clustering of the hits

For the clustering of the step-specific phenotypes, the R-implementation of the hierarchical method described by Ward (1963) has been applied using an Euclidean distance metric. Partitioning of the dendrogram into the clusters a–g was done manually. The number of protein–protein interactions according to the STRING database within the clusters was then compared with the number of protein–protein interactions within randomly assembled clusters of the same size, drawn from the same set of candidate genes. This step was repeated 300 times, the number of protein–protein interactions within clusters c (25) and d (11) always being higher than within the corresponding randomly assembled cluster (average 6.77 and 0.52, respectively; Jensen *et al*, 2009).

Supplementary information

Supplementary information is available at the *Molecular Systems Biology* website (www.nature.com/msb).

Acknowledgements

We thank Dr Klemens Rottner, Dr Axel Nohturfft, Dr Klaudia Giehl and members of the Hardt group for stimulating discussions, and Jacques Laville and the management of the Brutus cluster at ETH Zürich for excellent IT support. We also thank Dr Andreas Vonderheit for help with the MD microscope, Karin Mench and Dr Lilli Stergiou for help with the virus infections and Naomi Barrett and Silke Misselwitz for help with the figures. We are grateful to Dr Klaudia Ghiel for providing Rac1–GFP and CDC42–GFP constructs. BM was supported by a grant from the Bonizzi-Theler foundation. The project was supported by grants to WDH and BM from UBS AG on behalf of a customer and a grant (InfectX) to WDH from the Swiss SystemsX.ch initiative, evaluated by the Swiss National Science Foundation.

Author contributions: BM, SD, PV, LP and WDH designed the experiments; BM, SD, PV, RS, SR and MS performed the experiments; BM, SD, PV, BS, MS, CvM and WDH analyzed the results; and BM, SD, PV and WDH wrote the manuscript.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Agaisse H, Burrack LS, Philips JA, Rubin EJ, Perrimon N, Higgins DE (2005) Genome-wide RNAi screen for host factors required for intracellular bacterial infection. *Science* **309:** 1248–1251

Balcer HI, Goodman AL, Rodal AA, Smith E, Kugler J, Heuser JE, Goode BL (2003) Coordinated regulation of actin filament turnover by a high-molecular-weight Srv2/CAP complex, cofilin, profilin, and Aip1. *Curr Biol* **13:** 2159–2169

Barwe SP, Anilkumar G, Moon SY, Zheng Y, Whitelegge JP, Rajasekaran SA, Rajasekaran AK (2005) Novel role for Na,K-ATPase in phosphatidylinositol 3-kinase signaling and suppression of cell motility. *Mol Biol Cell* **16:** 1082–1094

Beck R, Rawet M, Wieland FT, Cassel D (2009) The COPI system: molecular mechanisms and function. *FEBS Lett* **583:** 2701–2709

Benjamini Y, Hochberg J (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B (Methodological)* **57:** 289–300

Bethune J, Wieland F, Moelleken J (2006) COPI-mediated transport. *J Membr Biol* **211:** 65–79

Carpenter AE, Jones TR, Lamprecht MR, Clarke C, Kang IH, Friman O, Guertin DA, Chang JH, Lindquist RA, Moffat J, Golland P, Sabatini DM (2006) CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol* **7:** R100

Charpentier X, Oswald E (2004) Identification of the secretion and translocation domain of the enteropathogenic and enterohemorrhagic *Escherichia coli* effector Cif, using TEM-1 beta-lactamase as a new fluorescence-based reporter. *J Bacteriol* **186:** 5486–5495

Chen LM, Hobbie S, Galan JE (1996) Requirement of CDC42 for *Salmonella*-induced cytoskeletal and nuclear responses. *Science* **274:** 2115–2118

Chen Y, Yang Z, Meng M, Zhao Y, Dong N, Yan H, Liu L, Ding M, Peng HB, Shao F (2009) Cullin mediates degradation of RhoA through evolutionarily conserved BTB adaptors to control actin cytoskeleton structure and cell movement. *Mol Cell* **35:** 841–855

Cheng LW, Viala JP, Stuurman N, Wiedemann U, Vale RD, Portnoy DA (2005) Use of RNA interference in *Drosophila* S2 cells to identify host pathways controlling compartmentalization of an intracellular pathogen. *Proc Natl Acad Sci USA* **102:** 13646–13651

Cherry S (2008) Genomic RNAi screening in *Drosophila* S2 cells: what have we learned about host-pathogen interactions? *Curr Opin Microbiol* **11:** 262–270

Chong R, Swiss R, Briones G, Stone KL, Gulcicek EE, Agaisse H (2009) Regulatory mimicry in Listeria monocytogenes actin-based motility. *Cell Host Microbe* **6:** 268–278

Coil DA, Miller AD (2004) Phosphatidylserine is not the cell surface receptor for vesicular stomatitis virus. *J Virol* **78:** 10920–10926

Conaway RC, Sato S, Tomomori-Sato C, Yao T, Conaway JW (2005) The mammalian mediator complex and its role in transcriptional regulation. *Trends Biochem Sci* **30:** 250–255

Cortes C, Vapnik V (1995) Support-vector networks. *AT&T Labs Res* **20:** 273–297

Dai S, Sarmiere PD, Wiggan O, Bamburg JR, Zhou D (2004) Efficient Salmonella entry requires activity cycles of host ADF and cofilin. *Cell Microbiol* **6:** 459–471

del Pozo MA, Alderson NB, Kiosses WB, Chiang HH, Anderson RG, Schwartz MA (2004) Integrins regulate Rac targeting by internalization of membrane domains. *Science* **303:** 839–842

Derre I, Pypaert M, Dautry-Varsat A, Agaisse H (2007) RNAi screen in *Drosophila* cells reveals the involvement of the Tom complex in Chlamydia infection. *PLoS Pathog* **3:** 1446–1458

Ehrbar K, Hapfelmeier S, Stecher B, Hardt WD (2004) InvB is required for type III-dependent secretion of SopA in Salmonella enterica serovar Typhimurium. *J Bacteriol* **186:** 1215–1219

Elwell CA, Ceesay A, Kim JH, Kalman D, Engel JN (2008) RNA interference screen identifies Abl kinase and PDGFR signaling in *Chlamydia trachomatis* entry. *PLoS Pathog* **4:** e1000021

Ewen SW, Naughton PJ, Grant G, Sojka M, Allen-Vercoe E, Bardocz S, Thorns CJ, Pusztai A (1997) Salmonella enterica var Typhimurium and Salmonella enterica var Enteritidis express type 1 fimbriae in the rat *in vivo*. *FEMS Immunol Med Microbiol* **18:** 185–192

Finlay BB, Ruschkowski S, Dedhar S (1991) Cytoskeletal rearrangements accompanying salmonella entry into epithelial cells. *J Cell Sci* **99** (Part 2)**:** 283–296

Friebel A, Ilchmann H, Aelpfelbacher M, Ehrbar K, Machleidt W, Hardt WD (2001) SopE and SopE2 from Salmonella Typhimurium activate different sets of RhoGTPases of the host cell. *J Biol Chem* **276:** 34035–34040

Gekara NO, Jacobs T, Chakraborty T, Weiss S (2005) The cholesterol-dependent cytolysin listeriolysin O aggregates rafts via oligomerization. *Cell Microbiol* **7:** 1345–1356

Goley ED, Welch MD (2006) The ARP2/3 complex: an actin nucleator comes of age. *Nat Rev Mol Cell Biol* **7:** 713–726

Guignot J, Caron E, Beuzon C, Bucci C, Kagan J, Roy C, Holden DW (2004) Microtubule motors control membrane dynamics of Salmonella-containing vacuoles. *J Cell Sci* **117:** 1033–1045

Hage B, Meinel K, Baum I, Giehl K, Menke A (2009) Rac1 activation inhibits E-cadherin-mediated adherens junctions via binding to IQGAP1 in pancreatic carcinoma cells. *Cell Commun Signal* **7:** 23

Hanisch J, Ehinger J, Ladwein M, Rohde M, Derivery E, Bosse T, Steffen A, Bumann D, Misselwitz B, Hardt WD, Gautreau A, Stradal TE, Rottner K (2010) Molecular dissection of Salmonella-induced membrane ruffling versus invasion. *Cell Microbiol* **12:** 84–98

Hao L, Sakurai A, Watanabe T, Sorensen E, Nidom CA, Newton MA, Ahlquist P, Kawaoka Y (2008) *Drosophila* RNAi screen identifies host genes important for influenza virus replication. *Nature* **454:** 890–893

Hapfelmeier S, Stecher B, Barthel M, Kremer M, Müller A, Heikenwalder M, Stallmach T, Hensel M, Pfeffer K, Akira S, Hardt WD (2005) The Salmonella Pathogenicity Island (SPI)-1 and SPI-2 Type III secretion systems allow Salmonella Serovar Typhimurium to trigger Colitis via MyD88-dependent and MyD88-independent mechanisms. *J Immunol* **174:** 1675–1685

Hardt WD, Chen LM, Schuebel KE, Bustelo XR, Galan JE (1998) *S. typhimurium* encodes an activator of Rho GTPases that induces membrane ruffling and nuclear responses in host cells. *Cell* **93:** 815–826

Hayward RD, Cain RJ, McGhie EJ, Phillips N, Garner MJ, Koronakis V (2005) Cholesterol binding by the bacterial type III translocon is essential for virulence effector delivery into mammalian cells. *Mol Microbiol* **56:** 590–603

Hirsch AJ (2010) The use of RNAi-based screens to identify host proteins involved in viral replication. *Future Microbiol* **5:** 303–311

Hoiseth SK, Stocker BA (1981) Aromatic-dependent *Salmonella typhimurium* are non-virulent and effective as live vaccines. *Nature* **291:** 238–239

Ikonen E (2008) Cellular cholesterol trafficking and compartmentalization. *Nat Rev Mol Cell Biol* **9:** 125–138

Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P, von Mering C (2009) STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res* **37:** D412–D416

Jutras I, Abrami L, Dautry-Varsat A (2003) Entry of the lymphogranuloma venereum strain of Chlamydia trachomatis into host cells involves cholesterol-rich membrane domains. *Infect Immun* **71:** 260–266

Kaniga K, Bossio JC, Galan JE (1994) The Salmonella Typhimurium invasion genes invF and invG encode homologues of the AraC and PulD family of proteins. *Mol Microbiol* **13:** 555–568

Kaplan JH (2002) Biochemistry of Na,K-ATPase. *Annu Rev Biochem* **71:** 511–535

Konig R, Stertz S, Zhou Y, Inoue A, Hoffmann HH, Bhattacharyya S, Alamares JG, Tscherne DM, Ortigoza MB, Liang Y, Gao Q, Andrews SE, Bandyopadhyay S, De Jesus P, Tu BP, Pache L, Shih C, Orth A, Bonamy G, Miraglia L *et al* (2010) Human host factors required for influenza virus replication. *Nature* **463:** 813–817

Kuijl C, Savage ND, Marsman M, Tuin AW, Janssen L, Egan DA, Ketema M, van den Nieuwendijk R, van den Eeden SJ, Geluk A, Poot A, van der Marel G, Beijersbergen RL, Overkleeft H, Ottenhoff TH, Neefjes J (2007) Intracellular bacterial growth is controlled by a kinase network around PKB/AKT1. *Nature* **450:** 725–730

Lafont F, Tran Van Nhieu G, Hanada K, Sansonetti P, van der Goot FG (2002) Initial steps of Shigella infection depend on the cholesterol/sphingolipid raft-mediated CD44-IpaB interaction. *EMBO J* **21:** 4449–4457

Lara-Tejero M, Galan JE (2009) The Salmonella Typhimurium SPI-1 type III secretion translocases mediate intimate attachment to non-phagocytic cells. *Infect Immun* **77:** 2635–2642

Lee H, Park DS, Razani B, Russell RG, Pestell RG, Lisanti MP (2002) Caveolin-1 mutations (P132L and null) and the pathogenesis of breast cancer: caveolin-1 (P132L) behaves in a dominant-negative manner and caveolin-1 (−/−) null mice show mammary epithelial cell hyperplasia. *Am J Pathol* **161:** 1357–1369

Maciver SK, Hussey PJ (2002) The ADF/cofilin family: actin-remodeling proteins. *Genome Biol* **3:** reviews3007

Marsman M, Jordens I, Kuijl C, Janssen L, Neefjes J (2004) Dynein-mediated vesicle transport controls intracellular Salmonella replication. *Mol Biol Cell* **15:** 2954–2964

McGhie EJ, Brawn LC, Hume PJ, Humphreys D, Koronakis V (2009) Salmonella takes control: effector-driven manipulation of the host. *Curr Opin Microbiol* **12:** 117–124

McGhie EJ, Hayward RD, Koronakis V (2004) Control of actin turnover by a salmonella invasion protein. *Mol Cell* **13:** 497–510

Medina E, Paglia P, Nikolaus T, Muller A, Hensel M, Guzman CA (1999) Pathogenicity island 2 mutants of Salmonella Typhimurium are efficient carriers for heterologous antigens and enable modulation of immune responses. *Infect Immun* **67:** 1093–1099

Meresse S, Steele-Mortimer O, Finlay BB, Gorvel JP (1999) The rab7 GTPase controls the maturation of Salmonella Typhimurium-containing vacuoles in HeLa cells. *EMBO J* **18:** 4394–4403

Misselwitz B, Kreibich SK, Rout S, Stecher B, Periaswamy B, Hardt WD (2011) Salmonella enterica serovar Typhimurium binds to HeLa cells via Fim-mediated reversible adhesion and irreversible type three secretion system 1-mediated docking. *Infect Immun* **79:** 330–341

Misselwitz B, Strittmatter G, Periaswamy B, Schlumberger MC, Rout S, Horvath P, Kozak K, Hardt WD (2010) Enhanced CellClassifier: a multi-class classification tool for microscopy images. *BMC Bioinform* **11:** 30

Muller AJ, Hoffmann C, Galle M, Van Den Broeke A, Heikenwalder M, Falter L, Misselwitz B, Kremer M, Beyaert R, Hardt WD (2009) The S. Typhimurium effector SopE induces caspase-1 activation in stromal cells to initiate gut inflammation. *Cell Host Microbe* **6:** 125–136

Munoz S, Rivas-Santiago B, Enciso JA (2009) Mycobacterium tuberculosis entry into mast cells through cholesterol-rich membrane microdomains. *Scand J Immunol* **70:** 256–263

Norris FA, Wilson MP, Wallis TS, Galyov EE, Majerus PW (1998) SopB, a protein required for virulence of *Salmonella dublin*, is an inositol phosphate phosphatase. *Proc Natl Acad Sci USA* **95:** 14057–14059

Orci L, Stamnes M, Ravazzola M, Amherdt M, Perrelet A, Sollner TH, Rothman JE (1997) Bidirectional transport by distinct populations of COPI-coated vesicles. *Cell* **90:** 335–349

Paavilainen VO, Bertling E, Falck S, Lappalainen P (2004) Regulation of cytoskeletal dynamics by actin-monomer-binding proteins. *Trends Cell Biol* **14:** 386–394

Palazzo AF, Eng CH, Schlaepfer DD, Marcantonio EE, Gundersen GG (2004) Localized stabilization of microtubules by integrin- and FAK-facilitated Rho signaling. *Science* **303:** 836–839

Patel JC, Galan JE (2005) Manipulation of the host actin cytoskeleton by Salmonella—all in the name of entry. *Curr Opin Microbiol* **8:** 10–15

Pelkmans L, Fava E, Grabner H, Hannus M, Habermann B, Krausz E, Zerial M (2005) Genome-wide analysis of human kinases in clathrin- and caveolae/raft-mediated endocytosis. *Nature* **436:** 78–86

Pepperkok R, Scheel J, Horstmann H, Hauri HP, Griffiths G, Kreis TE (1993) Beta-COP is essential for biosynthetic membrane transport from the endoplasmic reticulum to the Golgi complex *in vivo*. *Cell* **74:** 71–82

Petroski MD, Deshaies RJ (2005) Function and regulation of cullin-RING ubiquitin ligases. *Nat Rev Mol Cell Biol* **6:** 9–20

Philips JA, Rubin EJ, Perrimon N (2005) Drosophila RNAi screen reveals CD36 family member required for mycobacterial infection. *Science* **309:** 1251–1253

Pollard TD, Cooper JA (2009) Actin, a central player in cell shape and movement. *Science* **326:** 1208–1212

Popescu CI, Dubuisson J (2010) Role of lipid metabolism in hepatitis C virus assembly and entry. *Biol Cell* **102:** 63–74

Potrich C, Bastiani H, Colin DA, Huck S, Prevost G, Dalla Serra M (2009) The influence of membrane lipids in *Staphylococcus aureus* gamma-hemolysins pore formation. *J Membr Biol* **227:** 13–24

Prudencio M, Lehmann MJ (2009) Illuminating the host—how RNAi screens shed light on host-pathogen interactions. *Biotechnol J* **4:** 826–837

Pucadyil TJ, Schmid SL (2009) Conserved functions of membrane active GTPases in coated vesicle formation. *Science* **325:** 1217–1220

Ramet M, Manfruelli P, Pearson A, Mathey-Prevot B, Ezekowitz RA (2002) Functional genomic analysis of phagocytosis and

identification of a *Drosophila* receptor for *E. coli*. *Nature* **416:** 644–648

Ramo P, Sacher R, Snijder B, Begemann B, Pelkmans L (2009) CellClassifier: supervised learning of cellular phenotypes. *Bioinformatics* **25:** 3028–3030

Ridley AJ, Paterson HF, Johnston CL, Diekmann D, Hall A (1992) The small GTP-binding protein rac regulates growth factor-induced membrane ruffling. *Cell* **70:** 401–410

Riff JD, Callahan JW, Sherman PM (2005) Cholesterol-enriched membrane microdomains are required for inducing host cell cytoskeleton rearrangements in response to attaching-effacing *Escherichia coli*. *Infect Immun* **73:** 7113–7125

Rudolph MG, Weise C, Mirold S, Hillenbrand B, Bader B, Wittinghofer A, Hardt WD (1999) Biochemical analysis of SopE from *Salmonella typhimurium*, a highly efficient guanosine nucleotide exchange factor for RhoGTPases. *J Biol Chem* **274:** 30501–30509

Schlegel R, Tralka TS, Willingham MC, Pastan I (1983) Inhibition of VSV binding and infectivity by phosphatidylserine: is phosphatidylserine a VSV-binding site? *Cell* **32:** 639–646

Schlumberger MC, Hardt WD (2006) Salmonella type III secretion effectors: pulling the host cell's strings. *Curr Opin Microbiol* **9:** 46–54

Schlumberger MC, Kappeli R, Wetter M, Muller AJ, Misselwitz B, Dilling S, Kremer M, Hardt WD (2007) Two newly identified SipA domains (F1, F2) steer effector protein localization and contribute to Salmonella host cell manipulation. *Mol Microbiol* **65:** 741–760

Schlumberger MC, Muller AJ, Ehrbar K, Winnen B, Duss I, Stecher B, Hardt WD (2005) Real-time imaging of type III secretion: Salmonella SipA injection into host cells. *Proc Natl Acad Sci USA* **102:** 12548–12553

Seveau S, Bierne H, Giroux S, Prevost MC, Cossart P (2004) Role of lipid rafts in E-cadherin—and HGF-R/Met—mediated entry of Listeria monocytogenes into host cells. *J Cell Biol* **166:** 743–753

Shi J, Scita G, Casanova JE (2005) WAVE2 signaling mediates invasion of polarized epithelial cells by Salmonella Typhimurium. *J Biol Chem* **280:** 29849–29855

Shimaoka M, Springer TA (2003) Therapeutic antagonists and conformational regulation of integrin function. *Nat Rev Drug Discov* **2:** 703–716

Snijder B, Sacher R, Rämö P, Damm E, Liberali P, Pelkmans L (2009) Population context determines cell-to-cell variability in endocytosis and virus infection. *Nature* **461:** 520–523

Stecher B, Hapfelmeier S, Muller C, Kremer M, Stallmach T, Hardt WD (2004) Flagella and chemotaxis are required for efficient induction of Salmonella enterica serovar Typhimurium colitis in streptomycin-pretreated mice. *Infect Immun* **72:** 4138–4150

Styers ML, O'Connor AK, Grabski R, Cormet-Boyaka E, Sztul E (2008) Depletion of beta-COP reveals a role for COP-I in compartmentalization of secretory compartments and in biosynthetic transport of caveolin-1. *Am J Physiol Cell Physiol* **294:** C1485–C1498

Sun X, Yau VK, Briggs BJ, Whittaker GR (2005) Role of clathrin-mediated endocytosis during vesicular stomatitis virus entry into host cells. *Virology* **338:** 53–60

Tai AW, Benita Y, Peng LF, Kim SS, Sakamoto N, Xavier RJ, Chung RT (2009) A functional genomic screen identifies cellular cofactors of hepatitis C virus replication. *Cell Host Microbe* **5:** 298–307

Valderrama F, Luna A, Babia T, Martinez-Menarguez JA, Ballesta J, Barth H, Chaponnier C, Renau-Piqueras J, Egea G (2000) The golgi-associated COPI-coated buds and vesicles contain beta/gamma-actin. *Proc Natl Acad Sci USA* **97:** 1560–1565

van der Goot FG, Tran van Nhieu G, Allaoui A, Sansonetti P, Lafont F (2004) Rafts can trigger contact-mediated secretion of bacterial effectors via a lipid-based mechanism. *J Biol Chem* **279:** 47792–47798

Ward JH (1963) Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* **58:** 236–244

Welch MD, DePace AH, Verma S, Iwamatsu A, Mitchison TJ (1997) The human Arp2/3 complex is composed of evolutionarily conserved subunits and is localized to cellular regions of dynamic actin filament assembly. *J Cell Biol* **138:** 375–384

Wu WJ, Erickson JW, Lin R, Cerione RA (2000) The gamma-subunit of the coatomer complex binds Cdc42 to mediate transformation. *Nature* **405:** 800–804

Zhou D, Mooseker MS, Galan JE (1999) Role of the *S. typhimurium* actin-binding protein SipA in bacterial internalization. *Science* **283:** 2092–2095

References

[1] M. Stark, S. Berger, A. Stamatakis, and C. von Mering, "MLTreeMap - accurate maximum likelihood placement of environmental DNA sequences into taxonomic and functional reference phylogenies," *BMC Genomics*, vol. 11, no. 1, p. 461, 2010.

[2] C. von Linné, *Species plantarum, exhibentes plantas rite cognitas, ad genera relatas, cum differentiis specificis, nominibus trivialibus, synonymis selectis, locis natalibus, secundum systema sexuale digestas*. Lars Salvius, 1753.

[3] C. von Linné, *Systema naturæper regna tria naturæ, secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis*, vol. 1. Lars Salvius, 1758.

[4] B. Wiesemüller, H. Rothe, and W. Henke, *Phylogenetische Systematik: Eine Einführung*. Springer, Berlin, 1. ed., Sept. 2002.

[5] J. B. Reece, L. A. Urry, M. L. Cain, S. A. Wasserman, and P. V. Minorsky, *Campbell Biology*. Benjamin-Cummings Publishing Company, Subs of Addison Wesley Longman, Inc, 9. ed., Sept. 2010.

[6] E. Mayr, "Lamarck revisited," *Journal of the History of Biology*, vol. 5, no. 1, pp. 55–94, 1972.

[7] E. Jablonka and M. J. Lamb, "The changing concept of epigenetics," *Annals of the New York Academy of Sciences*, vol. 981, no. 1, pp. 82–96, 2006.

[8] A. A. Agrawal, C. Laforsch, and R. Tollrian, "Transgenerational in-

duction of defences in animals and plants," *Nature*, vol. 401, no. 6748, pp. 60–63, 1999.

[9] Z. A. Kaminsky, T. Tang, S. Wang, C. Ptak, G. H. T. Oh, A. H. C. Wong, L. A. Feldcamp, C. Virtanen, J. Halfvarson, C. Tysk, A. F. McRae, P. M. Visscher, G. W. Montgomery, I. I. Gottesman, N. G. Martin, and A. Petronis, "DNA methylation profiles in monozygotic and dizygotic twins," *Nat Genet*, vol. 41, pp. 240–245, Feb. 2009.

[10] A. O. Vargas, "Did paul kammerer discover epigenetic inheritance? A modern look at the controversial midwife toad experiments," *Journal of Experimental Zoology. Part B, Molecular and Developmental Evolution*, vol. 312, pp. 667–678, Nov. 2009. PMID: 19731234.

[11] E. R. Scerri, "The evolution of the periodic system," *Scientific American*, vol. 279, no. 3, pp. 56–61, 1998.

[12] S. G. Brush, "The reception of mendeleev's periodic law in america and britain," *Isis*, vol. 87, pp. 595–628, Dec. 1996.

[13] C. Stern, "The hardy-weinberg law," *Science*, vol. 97, pp. 137 –138, feb 1943.

[14] N. Eldredge, "Darwin's other books: 'red' and 'transmutation' notebooks, 'sketch,' 'essay,' and natural selection," *PLoS Biol*, vol. 3, p. e382, Nov. 2005.

[15] C. Darwin, *The Origin of Species by Means of Natural Selection: The Preservation of Favored Races in the Struggle for Life*. John Murray, Nov. 1859.

[16] J. Shoshani, C. P. Groves, E. L. Simons, and G. F. Gunnell, "Primate phylogeny: Morphological vs molecular results," *Molecular Phylogenetics and Evolution*, vol. 5, pp. 102–154, Feb. 1996.

[17] L. Kruckenhauser, E. Haring, W. Pinsker, M. J. Riesing, H. Winkler, M. Wink, and A. Gamauf, "Genetic vs. morphological differentiation of old world buzzards (genus buteo, accipitridae)," *Zoologica Scripta*, vol. 33, no. 3, pp. 197–211, 2004.

[18] D. M. Hillis, "Molecular versus morphological approaches to systematics," *Annual Review of Ecology and Systematics*, vol. 18, pp. 23–42, Jan. 1987.

[19] A. van Belkum, M. Struelens, A. de Visser, H. Verbrugh, and M. Tibayrenc, "Role of genomic typing in taxonomy, evolutionary genetics, and microbial epidemiology," *Clin. Microbiol. Rev.*, vol. 14, pp. 547–560, July 2001.

[20] M. L. Blaxter, "The promise of a DNA taxonomy," *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 359, pp. 669 –679, Apr. 2004.

[21] L. G. Wayne, D. J. Brenner, R. R. Colwell, P. A. D. Grimont, P. Kandler, M. I. Krichevsky, L. H. Moore, R. G. E. M. W. E. C. Moore, E. Stackebrandt, M. P. Star, and H. G. Truper, "Report of the ad hoc committee on reconciliation of approaches to bacterial systematics," *Int. J. Syst. Bacteriol.*, vol. 37, pp. 463–464, 1987.

[22] M. Blaxter, B. Elsworth, and J. Daub, "DNA taxonomy of a neglected animal phylum: an unexpected diversity of tardigrades," *Proceedings of the Royal Society B: Biological Sciences*, vol. 271, pp. S189–S192, 2004.

[23] E. Stackebrandt and J. Ebers, "Taxonomic parameters revisited: tarnished gold standards," *Microbiol Today*, vol. 8, no. 4, pp. 6–9, 2006.

[24] J. Goris, K. T. Konstantinidis, J. A. Klappenbach, T. Coenye, P. Vandamme, and J. M. Tiedje, "DNA-DNA hybridization values and their relationship to whole-genome sequence similarities," *Int J Syst Evol Microbiol*, vol. 57, pp. 81–91, Jan. 2007.

[25] K. T. Konstantinidis and J. M. Tiedje, "Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead," *Current Opinion in Microbiology*, vol. 10, pp. 504–509, Oct. 2007.

[26] E. Zuckerkandl, "Perspectives in molecular anthropology," *In Sherwood Larned Washburn (Ed.) Classification and human evolution*, pp. 243–272, 1964.

[27] E. Margoliash, "Primary structure and evolution of cytochrome c," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 50, pp. 672–679, Oct. 1963. PMID: 14077496 PMCID: 221244.

[28] W. M. Fitch and E. Margoliash, "Construction of phylogenetic trees,"

Science (New York, N.Y.), vol. 155, pp. 279–284, Jan. 1967. PMID: 5334057.

[29] C. R. Woese and G. E. Fox, "Phylogenetic structure of the prokaryotic domain: the primary kingdoms.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, pp. 5088–5090, Nov. 1977. PMID: 270744 PMCID: 432104.

[30] C. R. Woese, O. Kandler, and M. L. Wheelis, "Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 87, pp. 4576–4579, June 1990. PMID: 2112744.

[31] E. Suárez-Díaz and V. H. Anaya-Muñoz, "History, objectivity, and the construction of molecular phylogenies," *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, vol. 39, pp. 451–468, Dec. 2008.

[32] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, pp. 443–453, Mar. 1970. PMID: 5420325.

[33] J. L. Thorne, H. Kishino, and J. Felsenstein, "An evolutionary model for maximum likelihood alignment of DNA sequences," *Journal of Molecular Evolution*, vol. 33, pp. 114–124, Aug. 1991. PMID: 1920447.

[34] M. Brudno, S. Malde, A. Poliakov, C. B. Do, O. Couronne, I. Dubchak, and S. Batzoglou, "Glocal alignment: finding rearrangements during alignment," *Bioinformatics (Oxford, England)*, vol. 19 Suppl 1, pp. i54–62, 2003. PMID: 12855437.

[35] T. H. Jukes and C. R. Cantor, "Evolution of protein molecules," *In H. N. Munro (Ed.) Mammalian protein metabolism*, pp. 21–132, 1969.

[36] D. F. Feng, M. S. Johnson, and R. F. Doolittle, "Aligning amino acid sequences: Comparison of commonly used methods," *Journal of Molecular Evolution*, vol. 21, no. 2, pp. 112–125, 1985.

[37] T. Müller, R. Spang, and M. Vingron, "Estimating amino acid substitution models: A comparison of dayhoff's estimator, the resolvent

approach and a maximum likelihood method," *Molecular Biology and Evolution*, vol. 19, pp. 8 –13, Jan. 2002.

[38] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt, "A model of evolutionary change in proteins," *In M. O. Dayhoff & R. M. Schwartz (Eds.) Atlas of protein sequence and structure Vol. 5*, pp. 345–352, 1978.

[39] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 89, pp. 10915–10919, Nov. 1992. PMID: 1438297 PMCID: 50453.

[40] R. C. Edgar and S. Batzoglou, "Multiple sequence alignment," *Current Opinion in Structural Biology*, vol. 16, pp. 368–373, June 2006.

[41] L. Wang and T. Jiang, "On the complexity of multiple sequence alignment," *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, vol. 1, no. 4, pp. 337–348, 1994. PMID: 8790475.

[42] D. Feng and R. F. Doolittle, "Progressive sequence alignment as a prerequisitetto correct phylogenetic trees," *Journal of Molecular Evolution*, vol. 25, no. 4, pp. 351–360, 1987.

[43] M. Brudno, C. B. Do, G. M. Cooper, M. F. Kim, E. Davydov, NISC Comparative Sequencing Program, E. D. Green, A. Sidow, and S. Batzoglou, "LAGAN and Multi-LAGAN: efficient tools for Large-Scale multiple alignment of genomic DNA," *Genome Research*, vol. 13, pp. 721 –731, Apr. 2003.

[44] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Research*, vol. 32, pp. 1792 –1797, Mar. 2004.

[45] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins, "Clustal w and clustal x version 2.0," *Bioinformatics (Oxford, England)*, vol. 23, pp. 2947–2948, Nov. 2007. PMID: 17846036.

[46] J. H. Camin and R. R. Sokal, "A method for deducing branching sequences in phylogeny," *Evolution*, vol. 19, no. 3, pp. 311–326, 1965.

[47] R. A. Fisher, "On an absolute criterion for fitting frequency curves," *Messenger of Mathmatics*, vol. 41, pp. 155–160, 1912.

[48] J. Aldrich, "R.A. fisher and the making of maximum likelihood 1912-1922," *Statistical Science*, vol. 12, pp. 162–176, Sept. 1997.

[49] J. Felsenstein, "Cases in which parsimony or compatibility methods will be positively misleading," *Systematic Zoology*, vol. 27, pp. 401–410, Dec. 1978.

[50] J. W. Stiller and L. Harrell, "The largest subunit of RNA polymerase II from the glaucocystophyta: functional constraint and short-branch exclusion in deep eukaryotic phylogeny," *BMC Evolutionary Biology*, vol. 5, pp. 71–71, 2005. PMID: 16336687 PMCID: 1326215.

[51] F. S. L. Brinkman and D. D. Leipe, "Phylogenetic analysis," *In A. D. Baxenasis & B. F. F. Ouellete (Eds.) Bioinformatics: A practical guide to the analysis of genes and proteins (2nd ed.)*, 2001.

[52] J. Felsenstein, *Inferring Phylogenies*. Sinauer Associates, 2. ed., Sept. 2003.

[53] J. Felsenstein, "Evolutionary trees from DNA sequences: a maximum likelihood approach," *Journal of Molecular Evolution*, vol. 17, no. 6, pp. 368–376, 1981. PMID: 7288891.

[54] P. Lió and N. Goldman, "Models of molecular evolution and phylogeny," *Genome Research*, vol. 8, pp. 1233 –1244, Dec. 1998.

[55] J. P. Gogarten, H. Kibak, P. Dittrich, L. Taiz, E. J. Bowman, B. J. Bowman, M. F. Manolson, R. J. Poole, T. Date, and T. Oshima, "Evolution of the vacuolar H^+-atpase: implications for the origin of eukaryotes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 86, no. 17, pp. 6661 –6665, 1989.

[56] A. J. Roger, S. G. Svärd, J. Tovar, C. G. Clark, M. W. Smith, F. D. Gillin, and M. L. Sogin, "A mitochondrial-like chaperonin 60 gene in giardia lamblia: Evidence that diplomonads once harbored an endosymbiont related to the progenitor of?mitochondria," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, pp. 229 –234, Jan. 1998.

[57] R. J. Seviour, C. Kragelund, Y. Kong, K. Eales, J. L. Nielsen, and P. H. Nielsen, "Ecophysiology of the actinobacteria in activated sludge systems," *Antonie van Leeuwenhoek*, vol. 94, no. 1, pp. 21–33, 2008.

[58] R. W. Thacker and V. J. Paul, "Morphological, chemical, and genetic diversity of tropical marine cyanobacteria lyngbya spp. and symploca spp. (Oscillatoriales)," *Applied and Environmental Microbiology*, vol. 70, pp. 3305–3312, June 2004. PMID: 15184125.

[59] O. T. Avery, C. M. MacLeod, and M. McCarty, "Studies on the chemical nature of the substance inducing transformation of pneumococcal types," *The Journal of Experimental Medicine*, vol. 79, pp. 137 –158, Feb. 1944.

[60] R. M. Schwartz and M. O. Dayhoff, "Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts," *Science (New York, N.Y.)*, vol. 199, pp. 395–403, Jan. 1978. PMID: 202030.

[61] E. Hilario and J. P. Gogarten, "Horizontal transfer of ATPase genes–the tree of life becomes a net of life," *Bio Systems*, vol. 31, no. 2-3, pp. 111–119, 1993. PMID: 8155843.

[62] W. F. Doolittle, "Lateral genomics," *Trends in Cell Biology*, vol. 9, pp. M5–8, Dec. 1999. PMID: 10611671.

[63] W. F. Doolittle, "Phylogenetic classification and the universal tree," *Science (New York, N.Y.)*, vol. 284, pp. 2124–2129, June 1999. PMID: 10381871.

[64] W. F. Doolittle, "If the tree of life fell, would it make a sound?," *In J. Sapp (Ed.) Microbial phylogeny and evolution: Concepts and controversies*, pp. 119 –133, 2005.

[65] W. F. Doolittle, "Evolving biological organization," *In J. Sapp (Ed.) Microbial phylogeny and evolution: Concepts and controversies*, pp. 99 –118, 2005.

[66] F. Ge, L. Wang, and J. Kim, "The cobweb of life revealed by Genome-Scale estimates of horizontal gene transfer," *PLoS Biol*, vol. 3, no. 10, p. e316, 2005.

[67] M. Langille and F. Brinkman, "Bioinformatic detection of horizontally transferred DNA in bacterial genomes," *F1000 Biology Reports*, 2009.

[68] I. Choi and S. Kim, "Global extent of horizontal gene transfer," *Proceedings of the National Academy of Sciences*, vol. 104, pp. 4489 –4494, Mar. 2007.

[69] F. D. Ciccarelli, T. Doerks, C. von Mering, C. J. Creevey, B. Snel, and P. Bork, "Toward automatic reconstruction of a highly resolved tree of life," *Science*, vol. 311, pp. 1283 –1287, Mar. 2006.

[70] W. B. Whitman, D. C. Coleman, and W. J. Wiebe, "Prokaryotes: The unseen majority," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, pp. 6578 –6583, June 1998.

[71] M. T. Madigan, J. M. Martinko, P. V. Dunlap, and D. P. Clark, *Brock Biology of Microorganisms*. Boston: Pearson Benjamin-Cummings, 12. ed., 2008.

[72] R. Koch, "An address on bacteriological research," *British Medical Journal*, vol. 2, pp. 380–383, Aug. 1890. PMID: 20753110.

[73] R. de Wit and T. Bouvier, "'Everything is everywhere, but, the environment selects'; what did baas becking and beijerinck really say?," *Environmental Microbiology*, vol. 8, pp. 755–758, Apr. 2006. PMID: 16584487.

[74] F. Sanger, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, C. A. Fiddes, C. A. Hutchison, P. M. Slocombe, and M. Smith, "Nucleotide sequence of bacteriophage phi x174 DNA," *Nature*, vol. 265, pp. 687–695, Feb. 1977. PMID: 870828.

[75] R. I. Amann, W. Ludwig, and K. H. Schleifer, "Phylogenetic identification and in situ detection of individual microbial cells without cultivation," *Microbiological Reviews*, vol. 59, pp. 143–169, Mar. 1995. PMID: 7535888.

[76] M. S. Rappé and S. J. Giovannoni, "The uncultured microbial majority," *Annual Review of Microbiology*, vol. 57, pp. 369–394, 2003. PMID: 14527284.

[77] J. Xu, "Microbial ecology in the age of genomics and metagenomics: concepts, tools, and recent advances," *Molecular Ecology*, vol. 15, pp. 1713–1731, June 2006. PMID: 16689892.

[78] N. Dubilier, "The searchlight and the bucket of microbial ecology," *Environmental Microbiology*, vol. 9, no. 1, pp. 2–3, 2007.

[79] F. Rodríguez-Valera, "Environmental genomics, the big picture?," *FEMS Microbiology Letters*, vol. 231, pp. 153–158, Feb. 2004. PMID: 15027428.

[80] R. J. Ram, N. C. VerBerkmoes, M. P. Thelen, G. W. Tyson, B. J. Baker, R. C. Blake, M. Shah, R. L. Hettich, and J. F. Banfield, "Community proteomics of a natural microbial biofilm," *Science*, vol. 308, pp. 1915 –1920, June 2005.

[81] W. Liu and J. K. Jansson, *Environmental Molecular Microbiology*. Horizon Scientific Press, Jan. 2010.

[82] J. G. Bauman, J. Wiegant, P. Borst, and P. van Duijn, "A new method for fluorescence microscopical localization of specific DNA sequences by in situ hybridization of fluorochromelabelled RNA," *Experimental Cell Research*, vol. 128, pp. 485–490, Aug. 1980. PMID: 6157553.

[83] J. M. Levsky and R. H. Singer, "Fluorescence in situ hybridization: past, present and future," *J Cell Sci*, vol. 116, pp. 2833–2838, July 2003.

[84] E. Teira, T. Reinthaler, A. Pernthaler, J. Pernthaler, and G. J. Herndl, "Combining catalyzed reporter Deposition-Fluorescence in situ hybridization and microautoradiography to detect substrate utilization by bacteria and archaea in the deep ocean," *Applied and Environmental Microbiology*, vol. 70, pp. 4411–4414, July 2004. PMID: 15240332 PMCID: 444763.

[85] A. Pernthaler and J. Pernthaler, "Fluorescence in situ hybridization for the identification of environmental microbes," *Methods in Molecular Biology (Clifton, N.J.)*, vol. 353, pp. 153–164, 2007. PMID: 17332640.

[86] S. V. den Wyngaert, M. M. Salcher, J. Pernthaler, M. Zeder, and T. Posch, "Quantitative dominance of seasonally persistent filamentous cyanobacteria (Planktothrix rubescens) in the microbial assemblages of a temperate lake," *Limnology and Oceanography*, vol. 56, no. 1, pp. 97–109, 2011.

[87] V. J. Orphan, C. H. House, K. Hinrichs, K. D. McKeegan, and E. F. DeLong, "Methane-Consuming archaea revealed by directly coupled isotopic and phylogenetic analysis," *Science*, vol. 293, pp. 484 –487, July 2001.

[88] C. Lechene, F. Hillion, G. McMahon, D. Benson, A. M. Kleinfeld, J. P. Kampf, D. Distel, Y. Luyten, J. Bonventre, D. Hentschel, K. M. Park, S. Ito, M. Schwartz, G. Benichou, and G. Slodzian, "High-resolution quantitative imaging of mammalian and bacterial cells using stable isotope mass spectrometry," *Journal of Biology*, vol. 5, no. 6, pp. 20–20, 2006. PMID: 17010211 PMCID: 1781526.

[89] B. Glassner, C. Lechene, and Mcmahon, "Quantitative imaging of cells with multi-isotope imaging mass spectrometry (MIMS) - Nanoautography with stable isotope tracers," *Applied Surface Science*, vol. 252, no. 19, pp. 6895–6906, 2006.

[90] J. Chela-Flores, "Testing the universality of biology: A review," *International Journal of Astrobiology*, vol. 6, no. 03, pp. 241–248, 2007.

[91] N. R. Pace, D. A. Stahl, D. J. Lane, and G. J. Olsen, "Analyzing natural microbial populations by rrna sequences.," *ASM American Society for Microbiology News*, vol. 51, pp. 4–12, 1985.

[92] T. M. Schmidt, E. F. DeLong, and N. R. Pace, "Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing.," *Journal of Bacteriology*, vol. 173, pp. 4371–4378, July 1991. PMID: 2066334 PMCID: 208098.

[93] J. Handelsman, M. R. Rondon, S. F. Brady, J. Clardy, and R. M. Goodman, "Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products," *Chemistry & Biology*, vol. 5, pp. R245–249, Oct. 1998. PMID: 9818143.

[94] P. Lorenz, K. Liebeton, F. Niehaus, and J. Eck, "Screening for novel enzymes for biocatalytic processes: accessing the metagenome as a resource of novel functional sequence space," *Current Opinion in Biotechnology*, vol. 13, pp. 572–577, Dec. 2002.

[95] K. Liolios, I. A. Chen, K. Mavromatis, N. Tavernarakis, P. Hugenholtz, V. M. Markowitz, and N. C. Kyrpides, "The genomes on line database

(GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata," *Nucleic Acids Res.*, vol. 38, pp. D346–D354, Jan. 2010. PMID: 19914934 PMCID: 2808860.

[96] K. Zhang, A. C. Martiny, N. B. Reppas, K. W. Barry, J. Malek, S. W. Chisholm, and G. M. Church, "Sequencing genomes from single cells by polymerase cloning," *Nat Biotech*, vol. 24, pp. 680–686, June 2006.

[97] J. Handelsman, "Metagenomics: Application of genomics to uncultured microorganisms," *Microbiol. Mol. Biol. Rev.*, vol. 68, pp. 669–685, Dec. 2004.

[98] J. C. Venter, K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y. Rogers, and H. O. Smith, "Environmental genome shotgun sequencing of the sargasso sea," *Science (New York, N.Y.)*, vol. 304, pp. 66–74, Apr. 2004. PMID: 15001713.

[99] K. Mavromatis, N. Ivanova, K. Barry, H. Shapiro, E. Goltsman, A. C. McHardy, I. Rigoutsos, A. Salamov, F. Korzeniewski, M. Land, A. Lapidus, I. Grigoriev, P. Richardson, P. Hugenholtz, and N. C. Kyrpides, "Use of simulated data sets to evaluate the fidelity of metagenomic processing methods," *Nat Meth*, vol. 4, pp. 495–500, June 2007.

[100] J. Raes, J. O. Korbel, M. J. Lercher, C. von Mering, and P. Bork, "Prediction of effective genome size in metagenomic samples," *Genome Biology*, vol. 8, no. 1, p. R10, 2007. PMID: 17224063.

[101] F. E. Angly, D. Willner, A. Prieto-Davó, R. A. Edwards, R. Schmieder, R. Vega-Thurber, D. A. Antonopoulos, K. Barott, M. T. Cottrell, C. Desnues, E. A. Dinsdale, M. Furlan, M. Haynes, M. R. Henn, Y. Hu, D. L. Kirchman, T. McDole, J. D. McPherson, F. Meyer, R. M. Miller, E. Mundt, R. K. Naviaux, B. Rodriguez-Mueller, R. Stevens, L. Wegley, L. Zhang, B. Zhu, and F. Rohwer, "The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes," *PLoS Computational Biology*, vol. 5, p. e1000593, Dec. 2009. PMID: 20011103.

[102] P. L. F. Johnson and M. Slatkin, "Inference of microbial recombination rates from metagenomic data," *PLoS Genetics*, vol. 5, p. e1000674, Oct. 2009. PMID: 19798447.

[103] S. G. Tringe, C. von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, J. C. Detter, P. Bork, P. Hugenholtz, and E. M. Rubin, "Comparative metagenomics of microbial communities," *Science*, vol. 308, pp. 554 –557, Apr. 2005.

[104] J. A. Eisen, "Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes," *PLoS Biology*, vol. 5, p. e82, Mar. 2007. PMID: 17355177.

[105] H. Teeling, J. Waldmann, T. Lombardot, M. Bauer, and F. Glockner, "TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences," *BMC Bioinformatics*, vol. 5, no. 1, p. 163, 2004.

[106] R. Sandberg, G. Winberg, C. I. Bränden, A. Kaske, I. Ernberg, and J. Cöster, "Capturing whole-genome characteristics in short sequences using a naïve bayesian classifier," *Genome Research*, vol. 11, pp. 1404–1409, Aug. 2001. PMID: 11483581.

[107] A. C. McHardy, H. G. Martín, A. Tsirigos, P. Hugenholtz, and I. Rigoutsos, "Accurate phylogenetic classification of variable-length DNA fragments," *Nature Methods*, vol. 4, pp. 63–72, Jan. 2007. PMID: 17179938.

[108] H. Teeling, A. Meyerdierks, M. Bauer, R. Amann, and F. O. Glockner, "Application of tetranucleotide frequencies for the assignment of genomic fragments," *Environmental Microbiology*, vol. 6, no. 9, pp. 938–947, 2004.

[109] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F.

Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg, "Genome sequencing in microfabricated high-density picolitre reactors," *Nature*, vol. 437, no. 7057, pp. 376–380, 2005.

[110] H. Ji, N. Massé, S. Tyler, B. Liang, Y. Li, H. Merks, M. Graham, P. Sandstrom, and J. Brooks, "HIV drug resistance surveillance using pooled pyrosequencing," *PLoS ONE*, vol. 5, p. e9263, Feb. 2010.

[111] T. Abe, H. Sugawara, M. Kinouchi, S. Kanaya, and T. Ikemura, "Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples," *DNA Research*, vol. 12, pp. 281 –290, Jan. 2006.

[112] D. Kelley and S. Salzberg, "Clustering metagenomic sequences with interpolated markov models," *BMC Bioinformatics*, vol. 11, no. 1, p. 544, 2010.

[113] O. Nalbantoglu, S. Way, S. Hinrichs, and K. Sayood, "RAIphy: phylogenetic classification of metagenomics samples using iterative refinement of relative abundance index profiles," *BMC Bioinformatics*, vol. 12, no. 1, p. 41, 2011.

[114] D. Dalevi, D. Dubhashi, and M. Hermansson, "Bayesian classifiers for detecting HGT using fixed and variable order markov models of genomic signatures," *Bioinformatics (Oxford, England)*, vol. 22, pp. 517–522, Mar. 2006. PMID: 16403797.

[115] V. Kunin, A. Copeland, A. Lapidus, K. Mavromatis, and P. Hugenholtz, "A bioinformatician's guide to metagenomics," *Microbiol. Mol. Biol. Rev.*, vol. 72, pp. 557–578, Dec. 2008.

[116] D. H. Huson, A. F. Auch, J. Qi, and S. C. Schuster, "MEGAN analysis of metagenomic data," *Genome Research*, vol. 17, pp. 377–386, Mar. 2007. PMID: 17255551.

[117] P. Hugenholtz, B. M. Goebel, and N. R. Pace, "Impact of Culture-Independent studies on the emerging phylogenetic view of bacterial diversity," *J. Bacteriol.*, vol. 180, p. 6793, Dec. 1998.

[118] L. Krause, N. N. Diaz, A. Goesmann, S. Kelley, T. W. Nattkemper, F. Rohwer, R. A. Edwards, and J. Stoye, "Phylogenetic classification of short environmental DNA fragments," *Nucleic Acids Research*, vol. 36, pp. 2230 –2239, Apr. 2008.

[119] C. von Mering, P. Hugenholtz, J. Raes, S. G. Tringe, T. Doerks, L. J. Jensen, N. Ward, and P. Bork, "Quantitative phylogenetic assessment of microbial communities in diverse environments," *Science*, vol. 315, pp. 1126 –1130, Feb. 2007.

[120] M. Wu and J. A. Eisen, "A simple, fast, and accurate method of phylogenomic inference," *Genome Biology*, vol. 9, no. 10, p. R151, 2008. PMID: 18851752.

[121] J. Raes, K. U. Foerstner, and P. Bork, "Get the most out of your metagenome: computational analysis of environmental sequence data," *Current Opinion in Microbiology*, vol. 10, pp. 490–498, Oct. 2007. PMID: 17936679.

[122] E. Birney, M. Clamp, and R. Durbin, "GeneWise and genomewise," *Genome Research*, vol. 14, pp. 988–995, May 2004. PMID: 15123596.

[123] S. R. Eddy, "Profile hidden markov models," *Bioinformatics (Oxford, England)*, vol. 14, no. 9, pp. 755–763, 1998. PMID: 9918945.

[124] G. Talavera and J. Castresana, "Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments," *Systematic Biology*, vol. 56, pp. 564–577, Aug. 2007. PMID: 17654362.

[125] A. Stamatakis, "RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models," *Bioinformatics (Oxford, England)*, vol. 22, pp. 2688–2690, Nov. 2006. PMID: 16928733.

[126] N. Delmotte, C. Knief, S. Chaffron, G. Innerebner, B. Roschitzki, R. Schlapbach, C. von Mering, and J. A. Vorholt, "Community proteogenomics reveals insights into the physiology of phyllosphere bacteria," *Proceedings of the National Academy of Sciences*, vol. 106, no. 38, pp. 16428 –16433, 2009.

[127] V. Kunin, J. Raes, J. K. Harris, J. R. Spear, J. J. Walker, N. Ivanova, C. von Mering, B. M. Bebout, N. R. Pace, P. Bork, and P. Hugenholtz, "Millimeter-scale genetic gradients and community-level molecular convergence in a hypersaline microbial mat," *Molecular Systems Biology*, vol. 4, p. 198, 2008. PMID: 18523433.

[128] D. Wu, P. Hugenholtz, K. Mavromatis, R. Pukall, E. Dalin, N. N. Ivanova, V. Kunin, L. Goodwin, M. Wu, B. J. Tindall, S. D. Hooper, A. Pati, A. Lykidis, S. Spring, I. J. Anderson, P. D'haeseleer, A. Zemla, M. Singer, A. Lapidus, M. Nolan, A. Copeland, C. Han, F. Chen, J. Cheng, S. Lucas, C. Kerfeld, E. Lang, S. Gronow, P. Chain, D. Bruce, E. M. Rubin, N. C. Kyrpides, H. Klenk, and J. A. Eisen, "A phylogeny-driven genomic encyclopaedia of bacteria and archaea," *Nature*, vol. 462, pp. 1056–1060, Dec. 2009. PMID: 20033048.

[129] C. R. Woese, "Bacterial evolution," *Microbiological Reviews*, vol. 51, pp. 221–271, June 1987. PMID: 2439888.

[130] P. Yarza, W. Ludwig, J. Euzéby, R. Amann, K. Schleifer, F. O. Glöckner, and R. Rosselló-Móra, "Update of the All-Species living tree project based on 16S and 23S rRNA sequence analyses," *Systematic and Applied Microbiology*, vol. 33, pp. 291–299, Oct. 2010.

[131] D. J. Brenner, N. R. Krieg, J. T. Staley, and G. M. Garrity, eds., *Bergey's Manual of Systematic Bacteriology*. Boston, MA: Springer US, 2005.

[132] F. R. Tabita, T. E. Hanson, H. Li, S. Satagopan, J. Singh, and S. Chan, "Function, structure, and evolution of the RubisCO-Like proteins and their RubisCO homologs," *Microbiol. Mol. Biol. Rev.*, vol. 71, pp. 576–599, Dec. 2007.

[133] R. J. Ellis, "The most abundant protein in the world," *Trends in Biochemical Sciences*, vol. 4, pp. 241–244, Nov. 1979.

[134] H. Ashida, A. Danchin, and A. Yokota, "Was photosynthetic RuBisCO recruited by acquisitive evolution from RuBisCO-like proteins involved in sulfur metabolism?," *Research in Microbiology*, vol. 156, pp. 611–618, June 2005.

[135] L. J. Jensen, M. Kuhn, M. Stark, S. Chaffron, C. Creevey, J. Muller, T. Doerks, P. Julien, A. Roth, M. Simonovic, P. Bork, and C. von Mering, "STRING 8–a global view on proteins and their functional interactions in 630 organisms," *Nucleic Acids Research*, vol. 37, pp. D412–416, Jan. 2009. PMID: 18940858.

[136] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen, and C. von Mering, "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored," *Nucleic Acids Research*, vol. 39, pp. D561–568, Jan. 2011. PMID: 21045058.

[137] J. Raymond, J. L. Siefert, C. R. Staples, and R. E. Blankenship, "The natural history of nitrogen fixation," *Molecular Biology and Evolution*, vol. 21, pp. 541–554, Mar. 2004. PMID: 14694078.

[138] A. E. Dekas, R. S. Poretsky, and V. J. Orphan, "Deep-Sea archaea fix and share nitrogen in Methane-Consuming microbial consortia," *Science*, vol. 326, pp. 422 –426, Oct. 2009.

[139] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa, "KEGG for representation and analysis of molecular networks involving diseases and drugs," *Nucleic Acids Research*, vol. 38, pp. D355–360, Jan. 2010. PMID: 19880382.

[140] S. I. Chan, K. H. Chen, S. S. Yu, C. Chen, and S. S. Kuo, "Toward delineating the structure and function of the particulate methane monooxygenase from methanotrophic bacteria," *Biochemistry*, vol. 43, pp. 4421–4430, Apr. 2004. PMID: 15078087.

[141] C. Bedard and R. Knowles, "Physiology, biochemistry, and specific inhibitors of CH4, $NH4^+$, and CO oxidation by methanotrophs and nitrifiers.," *Microbiol. Mol. Biol. Rev.*, vol. 53, pp. 68–84, Mar. 1989.

[142] J. H. Rotthauwe, K. P. Witzel, and W. Liesack, "The ammonia monooxygenase structural gene amoA as a functional marker: molecular fine-scale analysis of natural ammonia-oxidizing populations.," *Applied and Environmental Microbiology*, vol. 63, pp. 4704–4712, Dec. 1997. PMID: 9406389 PMCID: 168793.

[143] T. Hayashi, R. Kaneko, M. Tanahashi, and T. Naganuma, "Molecular diversity of the genes encoding ammonia monooxygenase and particulate methane monooxygenase from deep-sea sediments," *Research Journal of Microbiology*, vol. 2, no. 6, pp. 530–537, 2007.

[144] M. Wagner, A. Loy, M. Klein, N. Lee, N. B. Ramsing, D. A. Stahl, and M. W. Friedrich, "Functional marker genes for identification of sulfate-reducing prokaryotes," *Methods in Enzymology*, vol. 397, pp. 469–489, 2005. PMID: 16260310.

[145] A. Loy, S. Duller, C. Baranyi, M. Mussmann, J. Ott, I. Sharon, O. Béjà, D. L. Paslier, C. Dahl, and M. Wagner, "Reverse dissimilatory sulfite reductase as phylogenetic marker for a subgroup of sulfur-oxidizing prokaryotes," *Environmental Microbiology*, vol. 11, pp. 289–299, Feb. 2009. PMID: 18826437.

[146] C. Lin and T. Todo, "The cryptochromes," *Genome Biology*, vol. 6, no. 5, pp. 220–220, 2005. PMID: 15892880 PMCID: 1175950.

[147] A. H. Singh, T. Doerks, I. Letunic, J. Raes, and P. Bork, "Discovering functional novelty in metagenomes: examples from light-mediated processes," *Journal of Bacteriology*, vol. 191, pp. 32–41, Jan. 2009. PMID: 18849420.

[148] V. M. Markowitz, E. Szeto, K. Palaniappan, Y. Grechkin, K. Chu, I. A. Chen, I. Dubchak, I. Anderson, A. Lykidis, K. Mavromatis, N. N. Ivanova, and N. C. Kyrpides, "The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions," *Nucleic Acids Research*, vol. 36, no. Database, pp. D528–D533, 2007.

[149] D. Gordon, C. Desmarais, and P. Green, "Automated finishing with autofinish," *Genome Research*, vol. 11, pp. 614–625, Apr. 2001. PMID: 11282977.

[150] K. A. Wetterstrand, "Dna sequencing costs: Data from the nhgri large-scale genome sequencing program." Available at: www.genome.gov/sequencingcosts. Accessed 2011.

[151] T. Junier and E. M. Zdobnov, "The newick utilities: high-throughput phylogenetic tree processing in the unix shell," *Bioinformatics*, vol. 26, pp. 1669 –1670, July 2010.

[152] P. N. Hess and C. A. D. M. Russo, "An empirical test of the midpoint rooting method," *Biological Journal of the Linnean Society*, vol. 92, no. 4, pp. 669–674, 2007.

[153] M. C. Schmid, A. B. Hooper, M. G. Klotz, D. Woebken, P. Lam, M. M. M. Kuypers, A. Pommerening-Roeser, H. J. M. O. den Camp, and M. S. M. Jetten, "Environmental detection of octahaem cytochrome c hydroxylamine/hydrazine oxidoreductase genes of aerobic and anaerobic ammonium-oxidizing bacteria," *Environmental Microbiology*, vol. 10, pp. 3140–3149, Nov. 2008. PMID: 18973625.

[154] M. Shimamura, T. Nishiyama, H. Shigetomo, T. Toyomoto, Y. Kawahara, K. Furukawa, and T. Fujii, "Isolation of a multiheme protein with features of a Hydrazine-Oxidizing enzyme from an anaerobic Ammonium-Oxidizing enrichment culture," *Applied and Environmental Microbiology*, vol. 73, pp. 1065–1072, Feb. 2007. PMID: 17172456 PMCID: 1828659.

[155] M. G. Klotz, M. C. Schmid, M. Strous, H. J. M. op den Camp, M. S. M. Jetten, and A. B. Hooper, "Evolution of an octahaem cytochrome c protein family that is key to aerobic and anaerobic ammonia oxidation by bacteria," *Environmental Microbiology*, vol. 10, pp. 3150–3163, Nov. 2008. PMID: 18761666.

[156] L. A. Sayavedra-Soto, N. G. Hommes, S. A. Russell, and D. J. Arp, "Induction of ammonia monooxygenase and hydroxylamine oxidoreductase mRNAs by ammonium in nitrosomonas europaea," *Molecular Microbiology*, vol. 20, pp. 541–548, May 1996. PMID: 8736533.

[157] M. Arumugam, E. D. Harrington, K. U. Foerstner, J. Raes, and P. Bork, "SmashCommunity: a metagenomic annotation and analysis tool," *Bioinformatics (Oxford, England)*, vol. 26, pp. 2977–2978, Dec. 2010. PMID: 20959381.